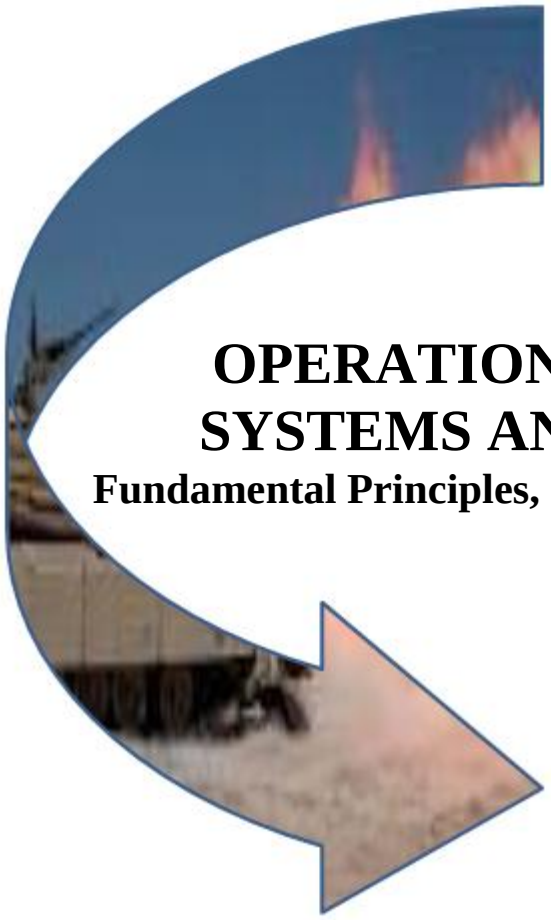


OPERATIONS RESEARCH/ SYSTEMS ANALYSIS (ORSA)

Fundamental Principles, Techniques, and Applications

A publication by:

The Army Logistics University
ORSA Committee
October 2011

Acknowledgement:

Thank you to each of the faculty of the ORSA committee of ALU that contributed in the writing and editing of this book as well as the creation of the excel templates.

For questions or suggestions about this publication contact:

The Army Logistics University
ORSA Committee
Email: leeeorsafam@conus.army.mil

TABLE OF CONTENTS

SECTION ONE
INTRODUCTION TO OPERATIONS RESEARCH
(Return to Table of Contents)

1. Purpose of this Guide. This reference guide is designed to provide basic information in the most common areas of Operations Research. These areas were selected based on their applicability to the types of problems most frequently encountered in our jobs and daily life. This guide will lead you through the basic techniques, processes and applications of Operations Research. It will take you through examples using conventional methods as well as templates created in Microsoft Excel.

2. Introduction.

   a. The twenty-first century has seen an explosion of technology and information. This has given us better ways of doing everything from transportation, housing, food production to defense. As individuals, we are constantly affected by these advances, both in the home and on the job.

   b. Along with these numerous advances has come an increasing awareness of the need to better control resources. We must learn to manage space, people, money, and materiel in such a way as to ensure the maximum benefit for our resource investments.

   c. Control of resources has become a daily task for most of us, both in our private lives and in our jobs, as demonstrated in Figure 1-1. To manage these resources we have developed a "library" of skills so to speak: tools which make it easier for us to control the resources for which we are responsible. These tools, referred to as management techniques, may be as simple as a checkbook ledger or as complex as a computerized inventory control system. But, they all have the same objective - better control of resources.

TYPE OF RESOURCE	AT HOME	AT WORK
Space	Domicile	Office, Warehouse, etc.
Personnel	Children's schedules	Subordinates, Peers and Supervisors
Capital	Paycheck	Revenue, Budget
Materiel	Groceries	Supplies

Figure 1-1. Resources we manage in our daily lives.

   d. Control of resources, however, involves more than just monitoring them. It also involves decisions on how to best allocate the resources available. Another set of management techniques is required to handle these decision-making problems. All of these techniques are grouped under what is most frequently referred to as Management Science or Operations Research.

3. History.

   a. Operations Research, or the application of scientific methods to management problems, is not a new concept. It has been employed in decision-making for many centuries. In fact, the earliest record of decision-making techniques dates back to ancient Babylonia (around 3200 B.C.). The consultants of the time employed quantitative ranking techniques in evaluating decision alternatives and recorded the results on day tablets.

   b. As far back as 500 B.C., the Chinese general, Sun Tzu, employed detailed analyses, including math formulas for logistics in the planning of his campaigns.

   c. During World War II, scientists were attached to the British and American military staffs to serve as scientific advisors on tactical and strategic problems. Among other things, they collected and analyzed statistical results from military operations. These analyses sometimes resulted in new tactical rules or in decisions regarding modernization of equipment and force structures. It was during this period that the term operations research was introduced, in reference to the application of scientific analysis to military operations. The following examples illustrate the use of the operations research approach:

      1) The British Coastal Command, in their aircraft attacks against submarines, changed from ordinary bombs that exploded on the surface of the water to depth charges. A depth of 100 feet was originally chosen, as the submarines could detect aircraft about two minutes before the attack and during this time could submerge about 100 feet. There was, however, a considerable uncertainty about this choice. The Coastal Command had, at this time, an operations research (OR) group under the leadership of Professor E. J. Williams. He started to collect statistical data and to analyze the chain of events during the attack. Numerical calculations showed that a change of the depth-setting to 25 feet should increase the chances of success of the average attack by a factor of two or more. Professor Williams succeeded in convincing the decision-making authorities and the depth-setting was changed. Within a few months of the change, the effectiveness of aircraft antisubmarine attacks was found to have increased as calculated.

      2) The studies of convoy tactics are among the most well known OR works from World War II. The work was initiated in connection with the presence of an OR analyst at a meeting at 10 Downing Street during the autumn of 1942. On this occasion, the best distribution of the limited shipyard capacity between merchant and escort vessels was being discussed. It was concluded that to make a comparison of the advantages of building more merchant ships or more escorts, the decrease in the number of merchant vessels sunk for each additional escort unit commissioned had to be known. A detailed analysis was started of the losses of vessels during the previous two years. The analysis showed that a convoy with nine escort vessels suffered an average of 25 percent fewer losses than a convoy with six escorts. On the basis of this result (and information about the number of convoys per year, the average size of convoys, and the number of existing escorts), it was calculated that every additional escort vessel would save two or

three merchant vessels a year. This led to the conclusion that more escorts should be built at the expense of the number of merchant vessels built.

3) Further analyses also showed that increased convoy speed, increased air escort protection and increased convoy size could decrease losses significantly. The last of these, the decrease in losses for larger convoys, turned out to be very important. The statistical data indicated that the percentage of vessels sunk from a larger convoy was smaller than the percentage of vessels sunk from a small convoy. This was contrary to the general opinion, according to which small convoys were considered comparatively safer. In order to investigate the correctness of this result, the OR group analyzed the different factors influencing the outcome of submarine attacks against convoys. On the basis of these studies a theoretical model was worked out which expressed the relation between convoy size, number of escort units and risk of sinking of the merchant vessel. The model verified and explained the unexpected finding. The Admiralty was convinced by the analysis, and new tactical rules prescribing increased convoy sizes were issued.

d. It was the application of scientific methods to strategic and tactical problems during World War II that cemented Operations Research as a formalized discipline. As a result of the effectiveness of operations research groups during WWII, industry as well as government took an active interest in continued application and development of Operations Research techniques.

e. The techniques of Operations Research (OR) have evolved along mathematical, statistical, numerical, and computer lines of theory and development (Figure 1-2).

Descriptive and Inferential Statistics	Differential & Integral Calculus	Regression and Correlation Analysis
Approximation Methods	Economic Analysis	Probability Theory
Artificial Intelligence	Forecasting	Queuing Theory
Cost Analysis	Game Theory	Network Analysis
Analysis of Variance	Math Programming	Simulation Techniques
Decision Analysis	Inventory Theory	Transportation Models
	Operational Effectiveness Analysis	

Figure 1-2. Operations Research Techniques

With the advent of advances in computer technology, the possibilities for successful use of different methods have substantially improved. Today, Operations Research techniques are employed widely throughout the Department of Defense and the U. S. Army. When asked the question, who does analysis? General Maxwell Thurman replied, “EVERYONE.” Whether determining what resources to use for a refueling mission, how to deploy units, or determining which combat systems to purchase in the future, soldiers to leaders are involved in the

analysis process. Among some of the more common applications are:

1) The application of statistical theory and sampling theory in systems development and testing.

2) The application of inventory theory and forecasting techniques in determining demand and establishing procurement levels, procurement quantities and lead times for major and secondary items.

3) Network analysis and transportation models for evaluating supply routes in operations plans and for project time and cost analysis.

4) Simulation and game theory in testing operational plans.

5) Cost Benefit Analysis in evaluation of competing systems.

4. Whether on the job or in their personal life, most people can benefit from Operations Research techniques. Decisions are made every day, applying the techniques discussed in this guide can provide you with the ability to evaluate those decisions in a more informed way.

SECTION TWO
THE STUDY PROCESS
(Return to Table of Contents)

1. Introduction. Commanders can no longer wholly rely on their prowess, wisdom and raw courage to conduct successful military operations. Battlefields are becoming increasingly complex, evolutionary changes in strategies and tactics are certain, and resource availability can be unpredictable. While experience and judgment have historically provided a strong basis for decision making, we also include more rational approaches in exploring force deployments, examining best strategies and tactics, and resource management. Analysis serves as means to that end.

2. The Study Process. Analytical study processes vary according to the task at hand. With each problem having its own set of peculiar requirements and special considerations, it is impractical to prescribe a step-by-step guide to address complex issues/problems. Analysts, therefore, must employ a scientific process (Figure 1). We can review a ten-step process across three primary phases – planning, preparation and execution. More specifically, the phases are commonly accepted as problem formulation, study methodology and communicating results. Each phase is essential, and therefore warrants further review.

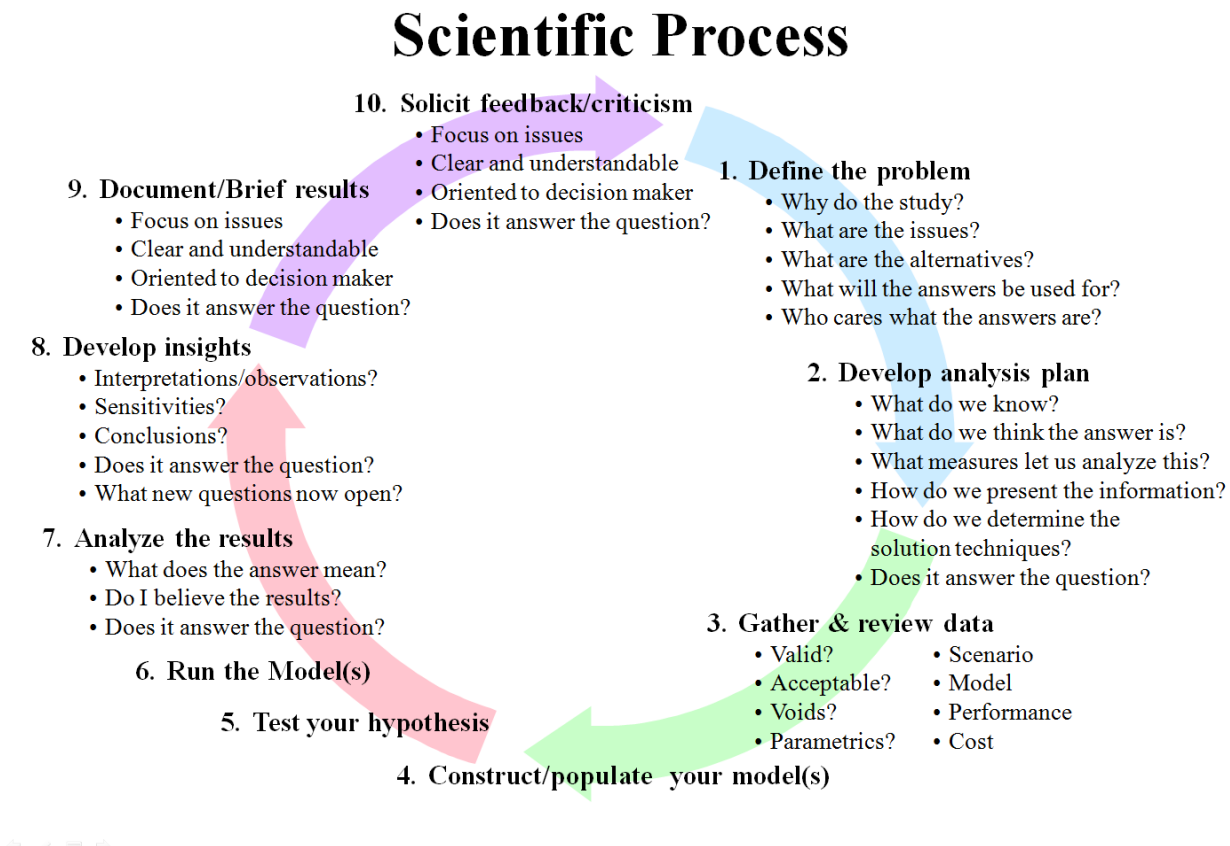


Figure 1

a. Problem Formulation. Problem Formulation begins the study process and requires substantial time and effort. More succinctly - it's just plain hard! But a problem well-stated is a problem half-solved. The formulation effort includes conducting background research, constructing a problem statement and study objectives, developing constraints, limitations and assumptions, identifying study issues and offering some preliminary analysis. A rudimentary methodology is well worth it.

(1) Conducting a background search, which includes analyzing the study directive, identifies the need for the study. It's more than simply reading the study directive, it requires coordination among stakeholders to ensure all grasp the nature of the study, its purpose, and expected deliverables.

(2) Conducting the background search can be seen as a fundamentally social process that fosters a shared understanding among all stakeholders. The effort generates a clear, concise, overarching and decomposable problem statement - from which all analysis originates. The study objective (there can be more than one) is a short, critical statement that describes what the study will accomplish.

(3) Developing constraints, limitations and assumptions organize the study. They set expectations for the study sponsor, frame the study space and set the stage for the methodology. How? Constraints serve as restrictions imposed by the study sponsor. They limit the study team's options in conducting the study. Limitations represent an inability – often breadth and depth – within the study team to fully meet the study objectives or fully investigate the study issues. Assumptions are educated guesses that replace facts that are not in evidence, but are important. Assumptions can quickly grow in number. Each, however, must be valid, must enable the study effort, is verifiable (sound and supportable) and necessary (generally agreed upon by all stakeholders); if not, it is quite possible to simply assume the problem away.

(4) Identifying study issues begins the process of decomposing the study problem into its constituent parts – in essence, creating smaller, more manageable study problems. These study issues, which are questions themselves, focus on and collectively answer the primary decision maker's problem. Such questions require answers. These answers, which come in the form of essential elements of analysis (EEA) and measures of merit (MoM), provide the crucial linkage between the problem, objective(s) and issues. EEAs, though questions too, provide answers which address the study objective(s). MoMs encompass different classes of measure; included are Measure of Effectiveness (MoE), Measure of Performance (MoP) and Data parameters (DP). While MoEs focus is on the impact of a system within its operational context, MoPs focus is on the internal system structure, characteristics, and behavior. DPs account for the properties or characteristics inherent in a system. Admittedly, it can be confusing. Figure 2 illustrates how a problem may be decomposed.

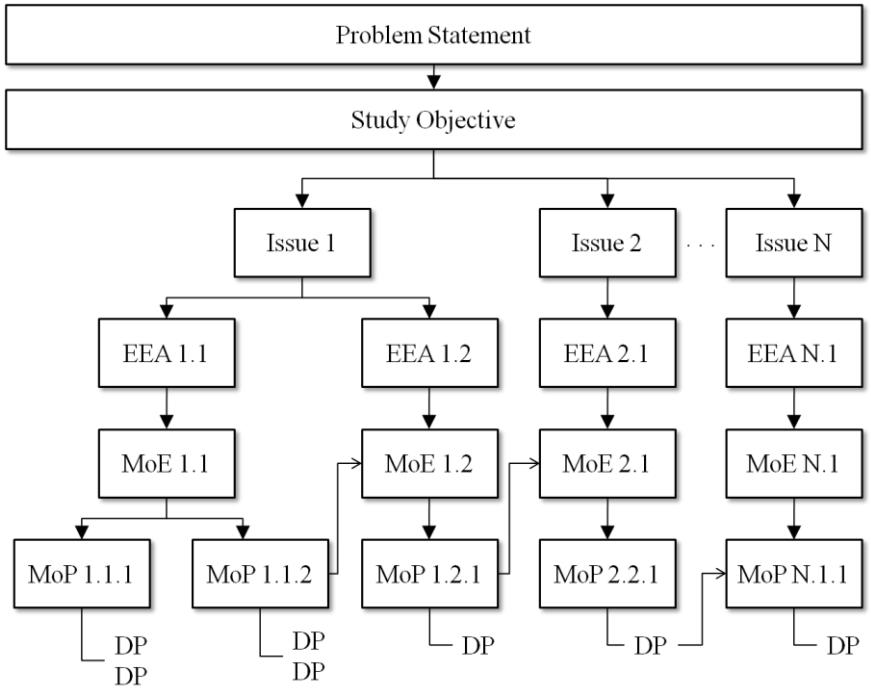


Figure 2

(5) Though time consuming, problem decomposition provides the opportunity to conduct preliminary analysis and construct a rudimentary study methodology. Consider the case when an essential element of analysis (which is a question) fails to adequately answer the associated issue. When the answer is yes, the issue is addressed; when the answer is no, the issue is unresolved and therefore fails to satisfy the study objective. In the first instance, a possible solution turns feasible; in the second instance, another possible solution becomes infeasible. The end result provides some preliminary analysis by way of reducing the study space with each infeasible solution. In constructing the initial study methodology, one needs to look no further than the measures of merit. Each such measure demands some qualitative or quantitative assessment, often derived from a tool or model of some fashion, which addresses the associated measure of effectiveness.

(6) To further the problem formulation process, consider the following example: Assume our study objective (stemming from an undisclosed problem statement) is *to destroy n targets using munitions x, y, and z*. Before any modeling can occur, one must identify the issue (or issues), EEAs, MoM and DPs. Figure 3 illustrates an example problem statement decomposition for the stated objective.

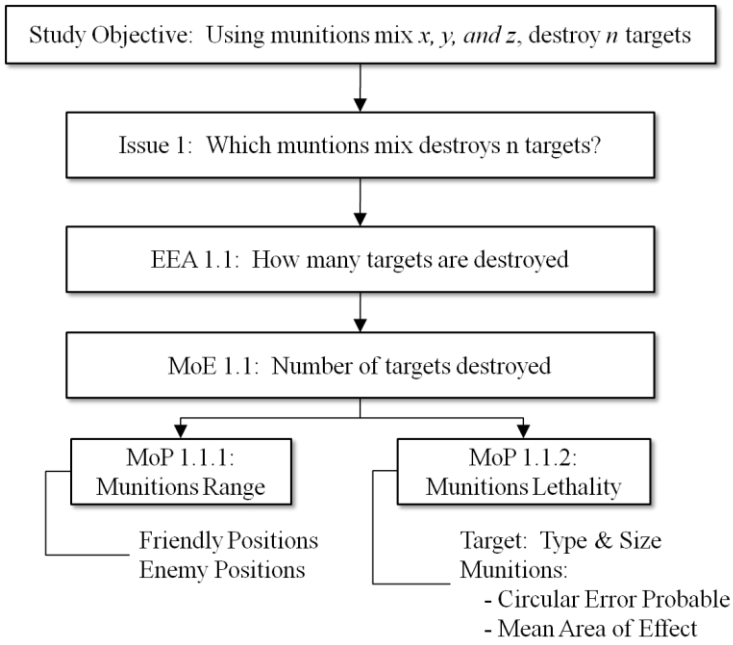


Figure 3

(7) Only when the problem formulation process fully addresses, rather than answers, the primary decision maker's issues can the study progress to the next study phase – study methodology.

b. Study Methodology. The study methodology provides the necessary linkage between problem formulation and the solution; it is the solution strategy, per se. The strategy, however, must operate within the confines of the study space – that is, within the bounds specifically prescribed by the problem formulation effort. The key to the linkage is properly selecting a model or models – or more exactly methods, tools and/or techniques – that are capable of meeting the study objective through fully addressing each of the measures of merit Figure 4 (top of next page).

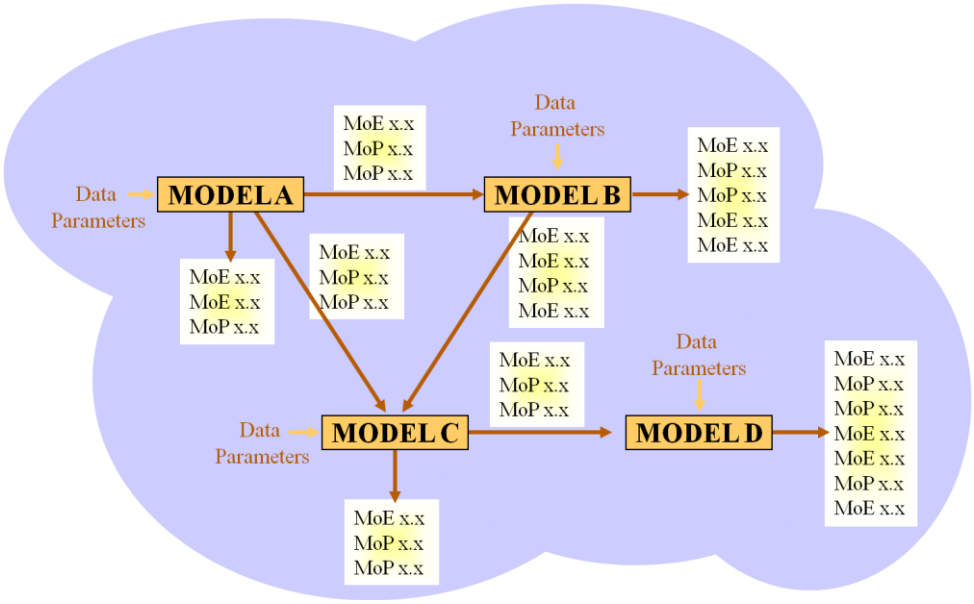


Figure 4

(1) What exactly are methods, tools and techniques? Methods suggest that a systematic (though sometimes very complex) approach to completing a task, or objective, exists. While tools serve as the mechanism toward accomplishing the task, techniques refer to the specific, and sometimes obvious, approaches toward efficiently completing it. Although seemingly connected vertically, the connections between methods, tools and techniques can span far – connections between models can also exist. These connections not only suggest that one model links directly to a measure of merit, but quite possibly to another model (i.e. where output from one model serves as input into another model). Although methods, tools and techniques afford effective model building, the modeling effort itself remains a challenge.

(2) What is modeling?

(a) Modeling is an art. Although there is no correct way to create art, DoD Publication 5000.59-P defines it as the application of a standard, rigorous, structured methodology to create and validate a model, which is the physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process. Otherwise stated, the modeling process requires development (conceptual), construction, implementation and, if necessary, modification.

[1] Model development begins the process (as noted in the problem formulation section), and is arguably the most important stage in the modeling process. The conceptual plan includes identifying the model's objectives, assumptions, input variables and output measures, as well as their relationships and interactions across system components.

[2] Model construction varies greatly; it can be simple or incredibly complex – but most importantly, it must be useful. Factors influencing model details include, but are not limited to the purpose, time and the nature of the phenomena. Is there need for the model to describe behavior or to prescribe a course of action? Is information known with certainty or is

there room for chance? Is there a need for the model to change over time? If there is need for such change, how does it occur? Such issues lend to the model of choice, which in these instances account for models that are in some combination descriptive or prescriptive, deterministic or probabilistic, static or dynamic, or measured discretely or continuously, respectively.

[3] Implementing the model is, in short, exercising it. Steps in this process include verification, validation and accreditation. While verification determines if the constructed model represents the conceptual model, validation determines if the constructed model is a reasonable representation of the real-world system. Accreditation occurs when the model can be applied for its intended purpose; it is often done by the developer, the decision maker or an external agency – with the latter giving it the greatest 'seal of approval'.

[4] Modification reflects the need for continual improvement, especially when new information on the phenomena becomes available.

(b) Again, models are reflective of the task at hand and therefore come in many sizes. Simple ones can often be solved with little rigor, yet still be effective. More complex models may require a full-on 'Study' effort. When employed properly, modeling results can be substantial.

(3) What exactly is meant by *properly employed*? Properly employed, quite simply, in this instance, is the ability to conclusively answer the measures of merit which, in turn, answers EEA(s) and Issue(s). When considering our study objective, the point can be made very clear: *to destroy n targets using munitions x,y, and z*. Through obtaining data parameter information on munitions characteristics (range, accuracy and damage capability) and target information (location, type and size), it is then possible to answer questions such as: can we range the target, can we hit the target and can we destroy the target. If one examines munitions range alone, it may be possible to reduce the number of possible munitions mixes. If one cannot range a target, then hitting and destroying a target with a particular munitions mix is a moot consideration. This then becomes an infeasible solution as MoE 1.1 & EEA 1.1 (which total to some value less than *n)*, fail to satisfactorily answer Issue 1 (Which munitions mix destroys *n* targets). On the other hand, if a particular munitions mix can range, hit and destroy all *n* targets then it is a viable consideration for further study.

(4) When considering further study on hitting and destroying targets, one can choose from myriad tools and techniques. The selection is often influenced by the resolution of the model, the availability of relevant data, and/or a combination of both. Let's put it together.

(a) A model's resolution can operate on many levels, but is commonly executed in either a low- or high-resolution mode; however, some more advanced models can accommodate a dynamic, multi-resolution mode. The choice, however, is a function of need. While low-resolution models aggregate the effect of phenomena, high-resolution models contain detailed information about all, or nearly all, phenomena of interest. For example, low-resolution models may employ a single (constant or variable) attrition rate to measure force-on-force effectiveness. High-resolution models assume a more disaggregated approach, that is,

attrition rates can be modeled (or paired) between individual munitions and any target. If we apply this to our example, we can say that:

- We can use a munitions mix (*x, y and z*) to attrite the opposing force (all *n* targets) and, - Attrition occurs at some fixed rate proportional to its strength/fighting capability - Then it suggests a low-resolution model.	- We can, for a single munitions types, relate its error probable (accuracy) to a particular target and, - We can determine the probability of a single shot hit and, - We can account for the number of successive munitions required to successfully hit that target and, - We determine the mean area of effect (ability to inflict damage) to a specific target to account for target destruction. - Then it suggests a high-resolution model.

Table 1

(b) Although high-resolution models provide a tremendous amount of battlefield detail (sometimes hundreds of thousands data points), these models can become unwieldy – so both low- and high-resolution models are actively used in modeling today…and both are capable of generating useful analysis.

(c) Relevant data can support analysis at any level in which it is available. Availability, however, is the key. When little to no data is available, over-arching assumptions (especially with force-on-force level attrition rates) may lend more towards employing a low-resolution model. In the absence of any data, one may be able to reasonably assume that one force may be able to attrite another at some rate (%) proportional to its relative strength. On the other hand, when ample, useful data exists, detailed analysis may occur – and a high-resolution model may be better suited for the task at hand. Such data may provide information that specifies that a single munitions type is 85% capable of achieving a single shot hit against one type of target, but only 30% capable against another type of target. In the end, it's a choice between applying relevant data and sound judgment.

(d) Compounding model resolution with available relevant data can often provide insight towards a solution methodology. Which models, tools and techniques are necessary? To what level of detail must each address? Are simple calculators enough? Is it necessary to construct a mathematical program? Can I? Would creating a simulation prove a better approach? Perhaps it is a combination of all of these ideas? There is no cookie-cutter approach; this is the point where science meets art!

(5) Methods, tools and techniques spearhead an analyst's approach in effectively addressing the study issue; each is essential in the solution strategy. Each provides some measure which helps in addressing the problem at hand. Problems are hardly tame, so model complexity increases quickly; there can be multiple, competing solutions. Balancing competing resources against constraints, limitations and assumptions can seem overwhelming. Figure 5. suggests one such complex model and the need for an optimizer to generate an optimal solution – in this case, a weapon-to-target pairing matrix of *munitions x, y and z against n targets*.

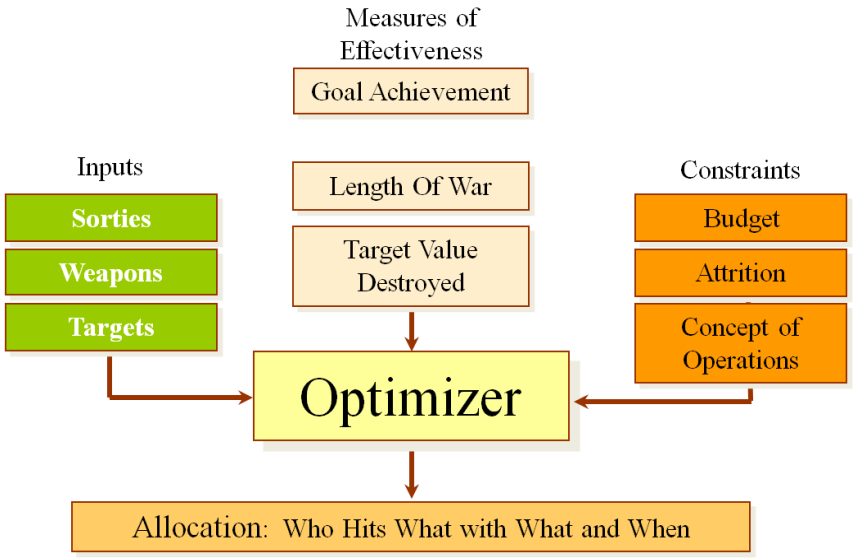


Figure 5

(6) Coupling integrated results, regardless of the solution methodology, with relevant sensitivity analysis provides the basis for findings, insights, and conclusions and/or recommendations.

(a) Sensitivity analysis illustrates to what extent the viability of a candidate solution (the product of integrated results) is influenced by perturbations across variables and/or parameters values. In essence, it provides answers to "what if" scenarios. Conducting the analysis is straightforward. Simply determine the key variables/parameters that may be sensitive, change the
variable's value (generally over a range of values) and then calculate the effect of the change against the candidate solution. While it is possible to also consider a combination of variable changes, multiple variable value changes can convolute results – consider changing only one variable at a time. In our example, rather than simply assuming a particular munitions ability to, in a single shot, hit and destroy a target to be a fixed value, say 85%, we could examine a range of values between 80% and 100%. What is the immediate resulting impact? What are the second and third order effects? Figure 6 depicts a monotonically increasing, nonlinear relationship between the number of rounds required to hit and destroy a target given a particular munitions ability (%). Considering the obvious trade-off, the 100% threshold might seem acceptable; it requires only five additional rounds (beyond the 95% threshold) to achieve hit and destroy certainty. If we consider the other effects, such as the munitions delivery method

and possible additional collateral damage, the additional 5% might prove too costly. When sensitivity analysis is conducted in a systematic fashion, interesting and often useful findings, insights and conclusions and/or recommendations emerge.

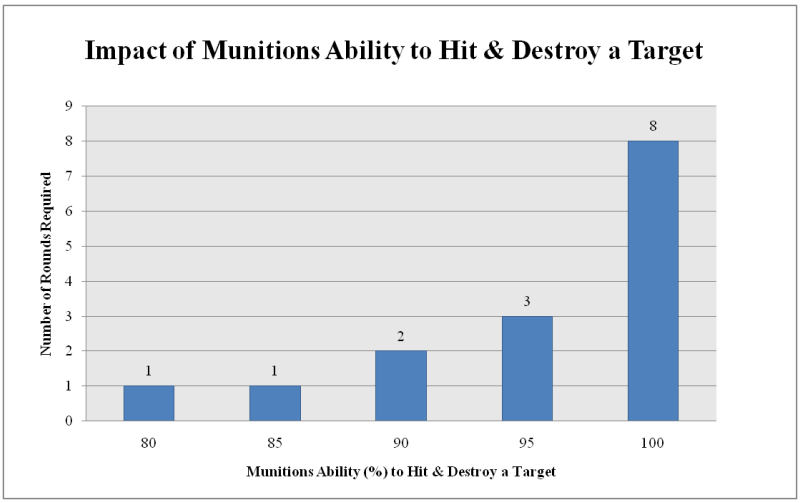


Figure 6

(b) Findings, insights and conclusions and/or recommendations provide the necessary information decision makers need to act. Distilling each from the analysis can be challenging, so let's consider their role and then an example. Findings reflect results. Insights identify relationships and behaviors (which can also suggest relative risks). Conclusions and/or recommendations reflect a reasoned deduction or inference – often referring to a potential solution. If applying these definitions to our example, then Table 2 may result.

Findings	With an 85% probability of a single shot hit, two or more rounds are required to achieve a stated probability of a kill greater than or equal to 90%.
Insights	Achieving a stated probability of a kill equal to 100% not only requires more munitions, but also extra delivery systems. The extra munitions negatively affect collateral damage nearly twofold. The extra delivery systems (for example, fixed wings) require more sorties, which may increase expected friendly casualties.
Conclusions and/or Recommendations	Accept 95% as the stated probability of a kill threshold which outweighs the cost associated with an additional 5% gain.

Table 2

c. Communicating Results.

(1) A formal, completed study plan requires translation to the responsible decision maker. Success is obtained when analytical results are put into operational terms that affect leadership decision making. The process is not an easy one; provided are a few tips.

(a) Ensure that your presentation accounts for not only the problem at-hand, but also the problem as it relates to other problems and/or situations. While there may be a single decision maker, the problem (and its solution) can extend to many others.

(b) Maintaining frequent, though scheduled, in-process reviews with key decision makers and/or stakeholders can serve as the impetus for sustaining the current study effort or the need for additional guidance. Additionally, in-process reviews are necessary when there are study plan changes, new key insights or unexpected developments occur. In-process reviews also serve to familiarize the study team with the expected audience receiving the study results.

(c) Presenting results can often be the only opportunity in which the analyst can influence the decision maker. Clear, concise and accurate results are often reflective of preparation. In the oral presentation, rehearsals are key. "Murder boards", which are practice sessions with peers, serve as an effective technique. External reviews can also serve equally, if not more, effective. In the written report, maintaining the report as an integral part throughout the study's problem formulation, methodology and results establishes and maintains the audit trail of thought. Preserving that development is essential. In either medium, seek out others that have completed a successful study, existing examples or obtain formal training accordingly.

(2) Effective communication occurs only when sponsor needs are fully addressed. More than the decision itself, it's the ability to provide the decision maker the ability to decide that earmarks successful translation.

6. Conclusion. The use of the Analysis Study Process is critical to ensuring that analysts provide sound analysis and alternatives that will enhance the military decision making process. The art of modeling and simulation has transformed military operations into an art-science, with trained, capable analysts serving as a fast-growing requirement for military commanders.

7. For additional information, consider the following references, from which much of this information was gleaned and/or transferred.

a. Study Directors' Guide: A Practical Handbook for Planning, Preparing, and Executing a Study (TRAC-F-TM-09-023)

b. Decision Making in Systems Engineering and Management, USMA, DSE

c. Field Manual (FM) 5.0

SECTON THREE
DESCRIPTIVE STATISTICS
(Return to Table of Contents)

1. Introduction.

a. We have heard statistics mentioned quite often. Some of the areas we hear statistics used are as follows:

baseball	football	weather report	advertisements
stock market	economic news	operational readiness	quality control

Statistics may be defined as:

The science of data which involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data to provide useful information.

b. Statistics is often divided into two fields:

(1) Descriptive Statistics is the branch of statistics devoted to collecting, organizing, summarizing, and describing entire data sets.

(2) Inferential Statistics is the branch of statistics concerned with using sample data to make an inference about a population.

c. We must first understand the groups from which data is collected. A population is a data set for the ENTIRE TARGET of our interest. For example if we are interested in the average miles driven by the army HMMWV fleet, the population would be the collection of all the miles driven by all the HMMWVs in the fleet. However, collecting data on an entire population is normally not feasible. It may involve too great a data set making it either time consuming or expensive to collect. Therefore, we usually collect a portion of data from the population. This is called a sample. A sample needs to have the following properties; it should be random, unbiased, and representative of the population.

d. The purpose of this section is to familiarize you with some of the descriptive and inferential techniques used in statistics and their computations.

2. Data. Data fall into two main categories; qualitative and quantitative and two measurement types; discrete and continuous.

a. Qualitative data. Qualitative data is data with no quantitative or mathematical interpretation. There are two categories of qualitative data.

(1) Nominal. This type of data is categorical data without any associated order of preference. Examples are town names, hair color, social security number, etc.

(2) Ordinal. This type of data is categorical data that has an order interpretation, but no true mathematical interpretation. Examples are hotel ratings; hotels rated 1 to 5 stars where more stars infer a better quality hotel; however a 4 star hotel does not indicate it is twice as good as a 2 star hotel. In the same manner placement in a race, 1^{st}, 2^{nd}, 3^{rd}, etc.

b. Quantitative data. Quantitative data is data that can be measured in a numerical scale. There are two categories of quantitative data will discuss.

(1) Interval. This type of data has an arbitrary zero with equal distance representing equal amounts. Examples are temperature, where 80 degrees Fahrenheit is 40 degrees warmer than 40 degrees Fahrenheit, however you cannot interpret that a it is twice as warm. That is due to the fact that 0 degrees Fahrenheit is not a true 0 measure as in no heat or cold.

(2) Ratio. This type of data has a meaningful zero value and all mathematical operations are valid. Examples are height, weight, time, etc.

c. Discrete data. Discrete data is considered countable. Examples are number of trucks in a motorpool, number of aircraft in a squadron, etc., where for example 6 ft. is twice as long as 3ft.

d. Continuous data. Continuous data is data that within any given interval there is an infinite amount of values. Examples are, height, weight, time.

3. Graphing data. After collecting data, it is always a good idea to graph the data. A pictorial view of the data can start describing the data you are evaluating as well as lead you in a direction of the analysis. Though there are many types of graphical representations, we will limit this discussion to two types, histograms and ogives. For this discussion we will use the following collection of test scores.

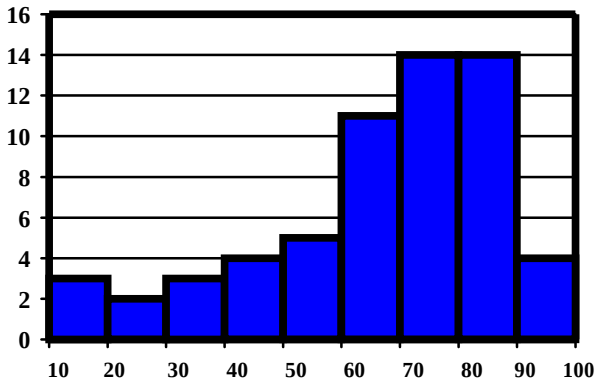
23	60	79	32	57	74	52	70	82	36
80	77	81	95	41	65	92	85	55	76
52	10	64	75	78	25	80	98	81	67
41	71	83	54	64	72	88	62	74	43
60	78	89	76	84	48	84	90	15	79
34	67	17	82	69	74	63	80	85	61

a. A frequency distribution shows the number of raw scores that have numerical values within each of a series of non-overlapping intervals. To construct the distribution, divide the range between 5-20 classes of equal width (more data, more classes). There are many techniques to determine how many classes to use. One simple technique is to use $\sqrt{n}$, where n is the number of values in the sample. This technique is useful as a guide, but you need to consider the data you are dealing with, and also the technique that will give you non- integer intervals. Once you have created the classes, place the data within an appropriate class and determine frequency.

True Limits	Frequency
10 and under 20	3
20 and under 30	2
30 and under 40	3
40 and under 50	4
50 and under 60	5
60 and under 70	11
70 and under 80	14
80 and under 90	14
90 and under 100	4
	60

Note: if we had used the $\sqrt{n}$ technique, we would have created 7 classes, however looking at the type of data, being test scores; they fit well into groupings of 10s.

b. A histogram takes the frequency distribution and creates a bar graph with it. To create a histogram, Mark the lower and upper limits of the class intervals on the horizontal axis. Then, Graph the frequency or the relative frequency on the vertical axis.

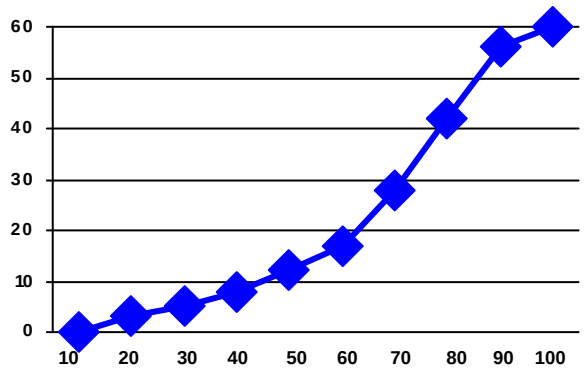


The histogram offers a graphical representation of the distribution of the test scores. We can note things such as the data are skewed left, the highest numbers of score are in the ranges between 60 and 90, and there are also many scores below 60. This type of information can be useful for a professor evaluating his class. An important feature of graphs is to lead you to be inquisitive; ask why and what. (i.e. Why are there so many scores below 60? What is the cause?, etc.)

c. A cumulative less than frequency is the total number of observations less than the upper limit of the interval.

Class boundaries	Frequency	Cumulative Less Than Frequency
Less than 20	3	3
Less than 30	2	5
Less than 40	3	8
Less than 50	4	12
Less than 60	5	17
Less than 70	11	28
Less than 80	14	42
Less than 90	14	56
Less than 100	4	60

d. An ogive is a linear representation of the cumulative less than data. To construct an ogive, mark the lower and upper limits of the class intervals on the horizontal axis, graph the cumulative less than frequency or the relative cumulative less than frequency above the upper limit of each class interval and graph 0 above the lower limit of the first interval.

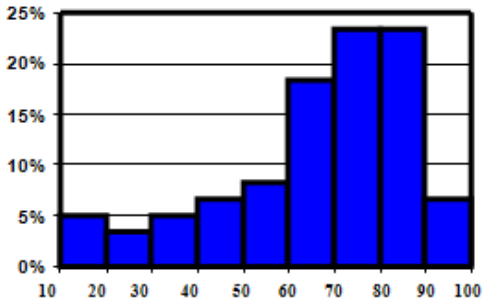


One feature of a graph such as this is the ability to detect the "knee in the curve"; that is, the point where there is a change in the inflection and slope. This point may be of some interest to investigate the reason for the change.

e. A relative frequency distribution is the proportion of the total number of observations that belong to each interval. To construct a relative frequency distribution, first construct a frequency distribution, then divide the frequency in each class by the total number of observations.

Class Interval	Frequency	Relative Frequency		
10 - 20	3	3/60 =	0.050 =	5.0%
20 - 30	2	2/60 =	0.033 =	3.3%
30 - 40	3	3/60 =	0.050 =	5.0%
40 - 50	4	4/60 =	0.067 =	6.7%
50 - 60	5	5/60 =	0.083 =	8.3%
60 - 70	11	11/60 =	0.183 =	18.3%
70 - 80	14	14/60 =	0.233 =	23.3%
80 - 90	14	14/60 =	0.233 =	23.3%
90 - 100	4	4/60 =	0.067 =	6.7%
		60/60 =	1.000 =	100.0%

f. A relative frequency histogram is constructed in the same manner as a frequency histogram except percentages are used on the vertical axis.

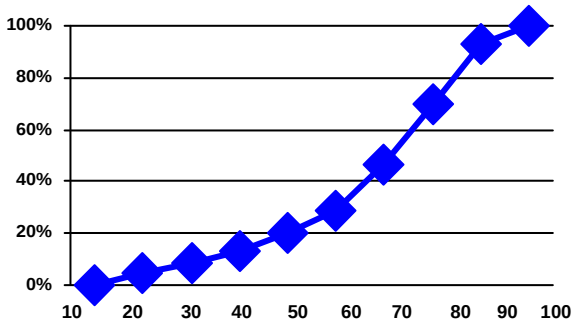


Note that this histogram takes on the same shape as the frequency histogram.

g. A relative cumulative less than frequency distribution is the proportion of the total number of observations less than the upper limit of the interval. To construct the interval, first construct the cumulative less than frequency distribution, then divide the cumulative frequency in each class by the total number of observations.

Class Boundaries	Relative Frequency	Cumulative Less Than Frequency
Less than 20	3	3/60 = 0.050 = 5.0%
Less than 30	2	5/60 = 0.083 = 8.3%
Less than 40	3	8/60 = 0.133 = 13.3%
Less than 50	4	12/60 = 0.200 = 20.0%
Less than 60	5	17/60 = 0.283 = 28.3%
Less than 70	11	28/60 = 0.467 = 46.7%
Less than 80	14	42/60 = 0.700 = 70.0%
Less than 90	14	56/60 = 0.933 = 93.3%
Less than 100	4	60/60 = 1.000 = 100.0%

h. An ogive is constructed for a relative cumulative less than frequency distribution.



Notice that this ogive is similar to the ogive on the last page – except that the Class Interval (0-60) has been replaced by cumulative percentages (0-100%).

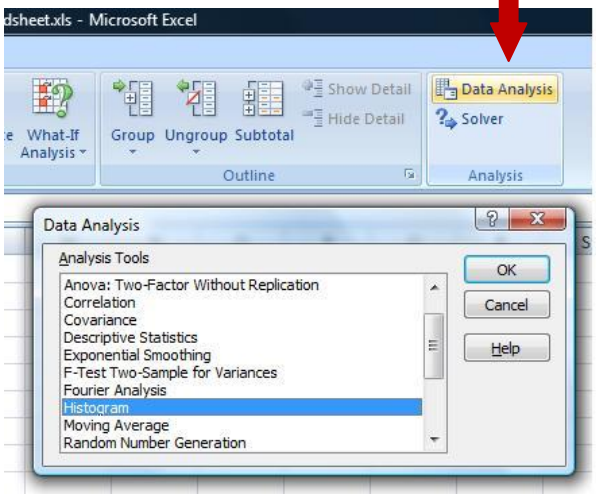
i. We'll now take a look at how we can graph this data using computers. We'll use Microsoft Excel 2007® to perform this function. Excel 2007® can use tools called add-ins to perform various functions. These programs use the features of Excel, but simplify them for the user by having drop down menus. Though you can purchase, or even create add-ins, Excel 2007® has several built in analysis tools. One such add-in is called **Data Analysis**. This add-in allows you to perform several basic statistical functions.

First, you will have to activate the Analysis ToolPak in Excel on your computer. To do this, go to **Appendix E** at the back of this manual. Appendix E will guide you through the steps to add the Analysis ToolPak to your Excel 2007® .

Once you have added the Analysis ToolPak to Excel, look at the image below. In Excel 2007 you will see a Data Tab at the top of your Excel Tab/Ribbon area. Click the Data Tab (left arrow below) and then look to the right side of the Data ribbon and you will see Data Analysis (right arrow below). Click the Data Analysis "button."

When you click the Data Analysis button you should see the Data Analysis menu appear (like the image on the right). Note that your menu will differ depending on the add-ins you have loaded.

We'll come back to this menu after we've entered some data. Click the **Cancel** button. Once we have entered the data we'll come back to the Data Analysis menu.

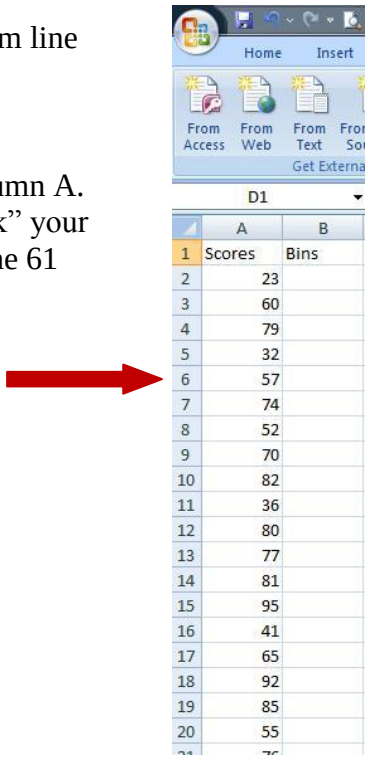


Now that you have Data Analysis activated, we can begin. First you'll add the data (the test scores discussed above) into Excel. On the right you'll see a partial image of the data that we initially entered into Excel.

First, we entered the word "Scores" into cell A1 and "Bins" into cell B1.

Then we entered the data, starting in cell A2 (down column A) from line one, then line two, and so on. See the image on the right.

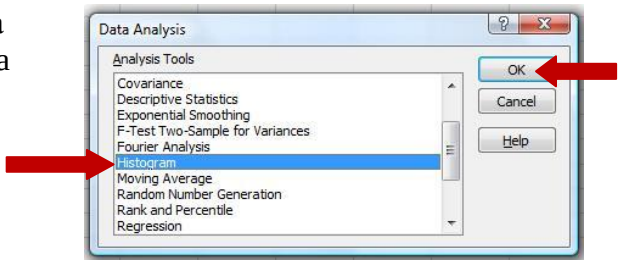
You can enter the numbers in any order/manner you desire in Column A. After you have entered the data, it would be wise to "double-check" your entries. There should sixty entries, and they should go down to line 61 (because we typed "Scores" into Cell A1.



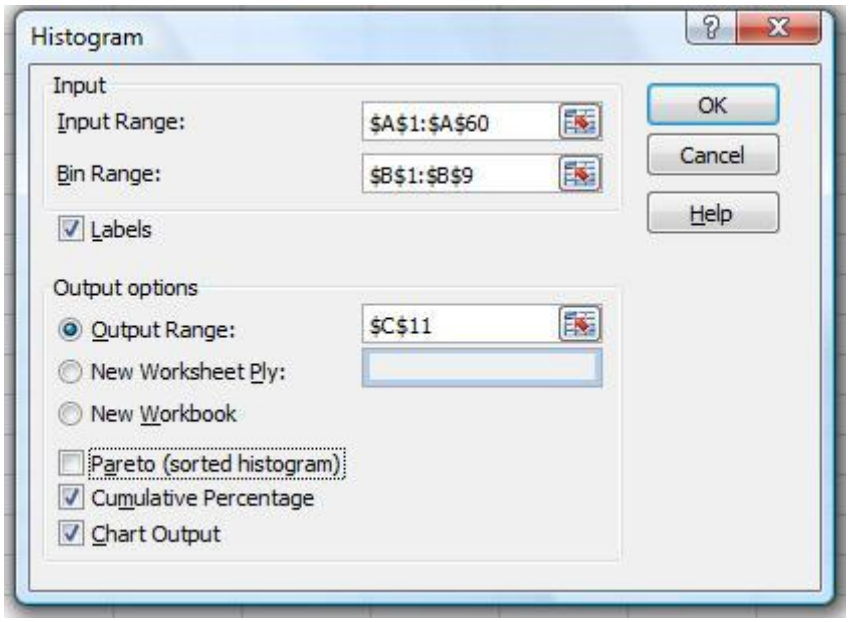
Next we'll add the upper limits (from Page 10) for each of the groups in Cells B2 – B9. The upper portion of your spreadsheet should look similar to the image on the right. Once you have the data, make sure it is in column form (like the image on the right and at the bottom of Page 13).

	A	B
1	Scores	Bins
2	23	20
3	60	30
4	79	40
5	32	50
6	57	60
7	74	70
8	52	80
9	70	90
10	82	

Now we'll go back to the Data Tab and Data Analysis button (top of Page 13). If the Data Tab is not selected, click the Data Tab, then click the Data Analysis button. Choose the **Histogram** tool then click the **OK** button.



A Histogram "Input Box" will now appear (similar to the one on the right).



If you are familiar with Excel, you may enter the data in any way you desire. If not, please type the data in each area – **just as you see it on the image on the right.**

First enter the data into the Input Range box. Then add the data for the Bin Range - called Bins on the spreadsheet. **If you added titles to the columns** (which we did) **- check the box marked labels** (by clicking in the box), this tells Excel the first row is not data.

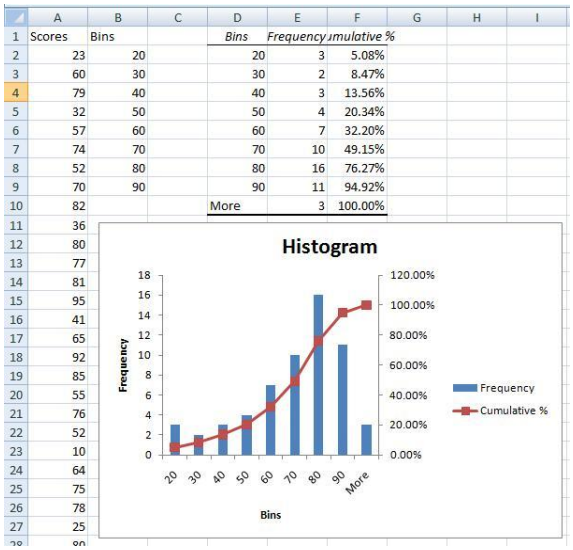
Next select where you want the graphic output to appear - Output Range. Notice we selected cell C1 ($C$1) – which is to the right of our data. Now select the type of graph you desire – we "checked" Cumulative Percentage and Chart Output. A chart output is a regular histogram; a cumulative percentage is a relative cumulative less than ogive. Your Histogram Input box should look like ours. When you are satisfied with the data - click the OK button.

Note that the program adds

- a frequency distribution and
- a relative cumulative less than frequency distribution

It also graphs the data in an embedded histogram and ogive.

You will have to adjust the graph and clean it to meet your presentation requirements.



	A	B	C	D	E	F	G	H	I
1	Scores	Bins		Bins	Frequency	Cumulative %			
2	23	20		20	3	5.08%			
3	60	30		30	2	8.47%			
4	79	40		40	3	13.56%			
5	32	50		50	4	20.34%			
6	57	60		60	7	32.20%			
7	74	70		70	10	49.15%			
8	52	80		80	16	76.27%			
9	70	90		90	11	94.92%			
10	82			More	3	100.00%			
11	36								
12	80								
13	77								
14	81								
15	95								
16	41								
17	65								
18	92								
19	85								
20	55								
21	76								
22	52								
23	10								
24	64								
25	75								
26	78								
27	25								

4. Measures of central tendency. These descriptive measures search for a value that is near the center of the data. We will discuss three measures; the mean, median, and mode.

a. The mean, also called the mathematical average, is given by adding all the values within a data set and then dividing by the total number of values within the set. The first formula is that for the mean of a **population**.

$\mu = \frac{\sum x}{N}$ The symbol μ, mu, is the Greek small letter "m" and represents a population mean. The symbol Σ, sigma, is the Greek capital "S" and is the summation symbol. This symbol tells us to add all the values that follow it. The x is a variable that represents each of the data values in our set. The N is the total number of values within the population. This formula then reads: Add all the values in the data set and divide by the total number of values

The next formula represents the mean of a sample.

$\bar{x} = \frac{\sum x}{n}$ We use the symbol $\bar{x}$, called x bar, to represent the sample mean. Note - the only other difference is that a small n is used to represent the total values within a sample.

Let's look at an example.

A sample of 10 orders is provided to the nearest pound.

The values are 3, 5, 4, 7, 5, 3, 4, 6, 8, 4.

Since this is a sample we use the second formula.

$$\bar{x} = \frac{3 + 5 + 4 + 7 + 5 + 3 + 4 + 6 + 8 + 4}{10} = \frac{49}{10} = 4.9$$

b. The median is that value which is in the center of the data set. To obtain the median for a data set, do the following:

(1) Arrange the data in ascending order.

(2) If the number of data points is *odd*, the median is the data point in the *middle*.

(3) If the number of data points is even, the median is the average of the two data points in the middle.

Looking at our example, we first arrange the data in order:

3, 3, 4, 4, 4, 5, 5, 6, 7, 8

Since we have an even number of data points, we take the two middle values, 4 and 5, and average them. That is add them together and divide by 2.

$$\text{median} = \frac{4+5}{2} = \frac{9}{2} = 4.5$$

c. The mode of a set of measurements is the value that occurs with the greatest Frequency (that is, *MOST OFTEN*). Data sets can have no mode if no values occur more than others, they can have one mode, two modes (bimodal), even three modes, but usually any more is said to have no mode. Looking at our data set, note that the value 4 occurs more than any other. Therefore, the mode = 4.

d. So what? When is it appropriate to use one type of measure versus another? We are very familiar with the use of the mean or average, however, the mean is skewed by extreme values. If you have a very high number, then the mean will tend to be high and may not represent your data set well. Therefore, in situations where these extreme points may skew the mean, the median is used. For example, we usually see this in income, where median income is reported vice average income, also in median housing costs. The mode is used, when you require knowledge of when most things occur, for example in retail, when do most customers shop? What sizes are used the most? What repair part breaks most often?

6. Measures of Dispersion. Let's consider the following two data sets:

A	B
1	46
10	49
50	50
50	50
90	51
99	54

Using the formulas we just learned, we will see that the mean, median, and mode for both data sets are 50. If you were to report only these measures to a person that had not actually seen the data sets, this person may believe both sets to be similar. However, you can easily observe that data set A has values that are more dispersed than data set B. Therefore, along with a measure of central tendency, we typically also report measures of dispersion. We will discuss three measures: range, variance, and standard deviation.

a. Range. The range is the difference between the largest and the smallest measurement in a data set, given by the formula.

Range = $x_{max} - x_{min}$

Therefore the range for data set A = 99 – 1 = 98.
And the range for data set B = 54 – 46 = 8.

These values give us a better picture of what the data sets look like. We definitely can see that data set A is more dispersed. However, the problem with only reporting the range is the misconception that the data points between the max and min values are evenly dispersed. In fact you could have a large range and only one value at an extreme point.

b. Variance. The variance is the average of the squared deviations from the mean. First let's look at the formula for the variance of a population.

$\boldsymbol{\sigma^2 = \frac{\sum(x-\mu)^2}{N}}$ We use the symbol $\boldsymbol{\sigma^2}$, sigma squared, which is the Greek small letter "s," to represent the population variance. Note that the formula tells us to subtract the mean from each value, then square each difference, add them together and finally divide by the total of the values in the population.

The sample variance is given by:

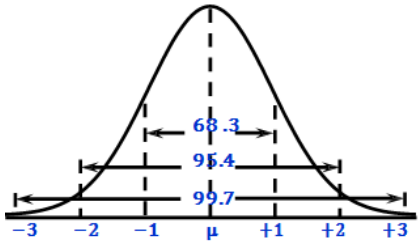
$\boldsymbol{s^2 = \frac{\sum(x-\bar{x})^2}{n-1}}$ The symbol $\boldsymbol{s^2}$, "s" squared, is used to represent sample variance. The only difference within the two formulas is that you divide by the total values in the sample minus one. This is done to account for the error found in using $\bar{x}$, which is an estimate of the population mean.

c. Standard Deviation. When we examine the formula of the variance we see that we squared the differences. In doing so, we also squared the units associated with them, i.e., squared feet, squared pounds. The standard deviation solves this by taking the positive square root of the variance and bringing the units back to it original form. The formulas are therefore:

$\sigma = \sqrt{\sigma^2}$ is the standard deviation for a population

$s = \sqrt{s^2}$ and s for a sample.

Lastly, if we look at the data in a normal distribution, we will see that a certain percent of the data fall within one, two, or three standard deviations plus or minus from the mean. Note that almost all the data falls within three standard deviations, hence the interest in what is known as "six sigma".



Let's continue the example of the 10 orders and find the range, variance, and standard deviation. The best method of doing this is to place the data into a chart and complete the computations.

A sample of 10 orders is provided to the nearest pound.

The values are: 3, 5, 4, 7, 5, 3, 4, 6, 8, 4.

Arrange the data in ascending order (from low to high). Find the mean. Subtract the mean from each value. Square each difference. Add the squared values. Divide that sum by n-1 to find the variance. Take the square root of the variance to find the standard deviation.

X	$\overline{X}$	$X - \overline{X}$	$(X - \overline{X})^2$
3	$4.\overline{9}$	-1.9	3.61
3	4.9	-1.9	3.61
4	4.9	-0.9	0.81
4	4.9	-0.9	0.81
4	4.9	-0.9	0.81
5	4.9	0.1	0.01
5	4.9	0.1	0.01
6	4.9	1.1	1.21
7	4.9	2.1	4.41
8	4.9	3.1	9.61
$\sum X = 49$			$\sum (X - \overline{X})^2 = 24.9$

Find the following measures.

1) sample mean $\frac{49}{10} = 4.9$

2) sample median $\frac{4+5}{2} = 4.5$

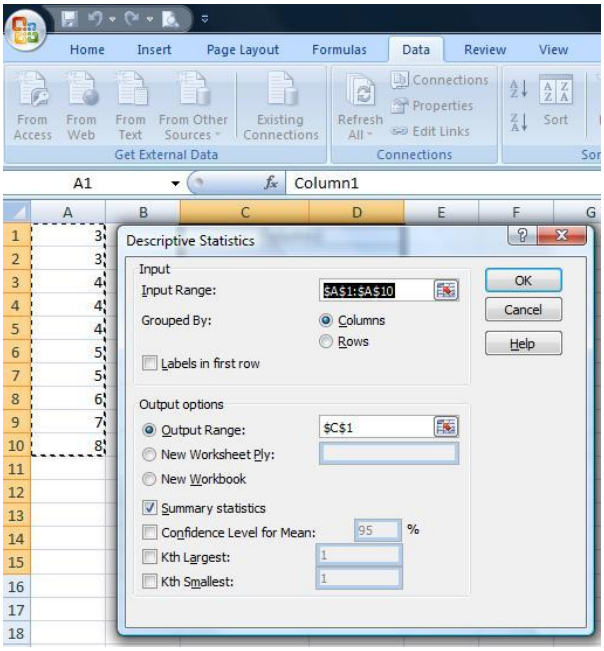
3) sample mode 4

4) sample range $8 - 3 = 5$

5) sample variance $\frac{24.9}{9} = 2.77$

6) sample standard deviation $\sqrt{2.77} = 1.66$

d. Let's go back to Excel 2007® and perform these descriptive measures using the Data Analysis add-in. Click the **Data Tab** like you did earlier. Then click the **Data Analysis button** - on the right side of the screen. This time, select **Descriptive Statistics**. Input the range of the data in the box and check labels in first row if appropriate. Select where you want the output and check summary statistics Your screen should look similar to the one on the right. Then click OK.



Note that all the measures we discussed are calculated to include a couple we did not. You must be aware that the program assumes the data is a sample and calculates based on that.

	A	B	C	D
1	3		*Column1*	
2	3			
3	4		Mean	4.9
4	4		Standard Error	0.52599113
5	4		Median	4.5
6	5		Mode	4
7	5		Standard Deviatic	1.66332999
8	6		Sample Variance	2.76666667
9	7		Kurtosis	-0.31043278
10	8		Skewness	0.73882854
11			Range	5
12			Minimum	3
13			Maximum	8
14			Sum	49
15			Count	10

SECTION FOUR
PROBABILITY AND DISTRIBUTIONS
(Return to Table of Contents)

1. Introduction.

   a. *What is probability?* Each of us has some idea of what that word means, but defining probability is difficult. Synonyms such as "chance," or "likelihood of occurrence," or "odds" may help us understand the concept, but do not actually define probability. Very loosely stated, a probability is a number that is used to represent or express uncertainty.

   b. The concept originated in the middle ages when mathematicians Pascal and Fermat attempted to describe gambling so that a winning strategy would be more apparent. This spawned ideas about risk and chance which are universally applicable.

   c. *How is probability determined?* The answer to this question depends upon the situation. One thing we can say about probability is that it has a definite range. A probability cannot be negative, nor can it exceed 1. That is, all values of probability must lie between 0 and 1, inclusive.

   d. Probabilities are used in everyday life:
      1) The percent chance of rain.
      2) The odds of your favorite team winning the Super Bowl.
      3) The chances of an auto accident which may be used to determine auto insurance rates.

   e. Probabilities are also used in a wide variety of analytic models:
      1) The probability of kill of a certain weapon system against another in a given scenario.
      2) The probability of a soldier passing a skills qualification test.
      3) The probability of a cost estimate being within a specified range.

   f. In all cases, the probabilities represent an attitude or a perception concerning the likelihood of an event occurring. The higher the probability, the more likely a particular event will occur. Given the fact that no circumstances from our point of view are certain, it is easy to see why probability is such a critical component of any type of problem. The old adage of "The only things certain in life are death and taxes!" has some merit but is not a true probability situation since it is considered an absolute; they will happen. The question is not if, but when they will happen.

g. Perceived certainty is actually a high probability of occurrence (i.e. higher than .90 or .95). For example: Is it a certainty that the sun will rise in the morning or is the idea based on historical data? Astronomers' knowledge as to how much longer the sun will continue to shine is speculative at best. For all we know, it could blow up tomorrow. So the conclusion is based on historical data: it has always risen, so it always will!

2. Probability Interpretations.

a. There are several interpretations of probability. One interpretation states that the probability of an event occurring is calculated by dividing the number of outcomes on which the event occurs by the total number of possible outcomes.

$$\mathbf{P(Event)} = \frac{\textbf{Successful Outcomes}}{\textbf{Total Possible Outcomes}}$$

This interpretation assumes that all outcomes are equally likely. Thus, the probability of rolling a 7 with a pair of dice is 6 favorable outcomes divided by 36 possible outcomes, which equals 1/6. This is called the classical or objective interpretation of probability.

b. In many situations the outcomes are not equally likely and the classical interpretation is of little or no value. It may be possible to use historical data in such situations to assess the probability. The probability of an event can be determined by the ratio of the frequency of favorable outcomes in a particular situation to the total number of outcomes.

$$\mathbf{P(Event)} = \frac{\textbf{Number of Times the Event Occurred}}{\textbf{Total Number of Times the Experiment was Repeated}}$$

This is the relative frequency probability concept. An example of this would be determining the probability of a soldier passing a particular skills qualification test by dividing the number of soldiers who passed the test by the number of soldiers who took it. Two key questions need to be considered when using a relative frequency interpretation:

1) Does the historical information accurately represent the current situation?
2) How many trials (pieces of historical information) are needed to accurately measure the probability?

c. When we consider the outcome of an experiment which is not repeatable, however, the relative frequency concept is of no use to us. An example is: "What is the probability that the cost of building a certain missile will not exceed $2 million?" When we come to such a situation, we are at a loss as to how to assess or calculate such a probability. One might wonder if the question even has any meaning. Certainly any answer to it is opinionated. This is the third interpretation of probability that is called subjective or personal probability.

d. One of the main complaints generally raised about the subjective approach to probability is simply that it is subjective. Even experts may well come up with radically different probability estimates. This is not a weakness of either classical or relative frequency probability. However, supporters of subjective probability are eager to point out that subjective estimates take into account qualitative as well as quantitative considerations, intangible as well as tangible points, and feelings as well as observations. Subjective probability is used quite often because the alternative is often using nothing.

e. It should be apparent by now that probability is a rather involved concept. Thus, when someone speaks in probabilistic terms, we should ask what probability concept he is using: classical, experimental, subjective, or perhaps a combination of several of these.

3. Probability Definitions.

a. Random phenomena: Any observed behavior of man, method, machine, material, or environment.

b. Experiment: Any process used to observe, count or measure random phenomena. For example, one experiment might be to fire a number of missiles and record the number of times they hit the target. Other experiments might be to roll one die and record the results or to select a group of Army officers and record the schools from which they received their undergraduate degrees.

c. Sample point: A single possible outcome of an experiment. In the die example, a roll of two would be considered a sample point. In the officer example, if a West Point graduate is chosen, then West Point would be a sample point.

d. Sample space: The set of all possible outcomes (sample points) of an experiment. For the die example, the sample space, which we will call S, would be: S = {1,2,3,4,5,6}. In the officer example, the sample space would be all of the various schools where the officers obtained degrees. The sum of all the probabilities in the sample space equals 1, P(S) = 1. We can therefore also conclude that P(null set) = 0.

e. Event: The desired outcome or set of outcomes. In probability an event is usually identified by a capital letter. If event A in the die problem is defined as the occurrence of an odd number, then A = {1,3,5}. If event B in the officer problem is defined as selecting a West Point officer, then event B would be the set of all officers who graduated from West Point. P(A) = ½ is read as the probability of event A equals ½.

f. Complement of an event: The complement of event A is the event that A *does not occur*. The complement of A would include all of the outcomes in the sample space for which A does not occur. The following notation can be used to represent a complement: A', A^c, or $\overline{A}$. In the dice problem if A = {1,3,5}, then $\overline{A}$ = {2,4, 6}. The probability of an event and its complement must sum to one or $P(A) + P(\overline{A}) = 1$.

g. Union of events: The set containing all the outcomes belonging to at least one of the events in the union. In the die experiment if event B = {1,2,3} and C = {3,4,5}, then the union of events B and C would contain the set of outcomes {1,2,3,4,5}. The notation for the union of B and C would be (B or C) or $(B \cup C)$, where $\cup$ indicates union. $(B \cup C) = \{1,2,3,4,5\}$.

h. Mutually exclusive events: When no two events have any sample points in common. Mutually exclusive events cannot occur together. In our die problem, the events of rolling a two and rolling an odd number are mutually exclusive as it is impossible to obtain both a two and an odd number on a single roll of a die. The events of getting a degree from West Point and a degree from Georgia Tech are mutually exclusive (unless someone has multiple degrees). Therefore, the probability of the union of a set of mutually exclusive events is equal to the sum of their individual probabilities. In the officer example the probability of selecting an officer with a degree from West Point or an officer with a degree from Georgia Tech would be equal to the sum of the probability of selecting an officer with a West Point degree and the probability of selecting an officer with a Georgia Tech degree.

i. Collectively exhaustive events: When all of the possible outcomes of the sample space are included in the set of defined events. In the die problem, the events of rolling an odd number and rolling an even number are collectively exhaustive (as well as mutually exclusive). If all of the schools that officers attended were represented in a set of events, then that set of events would be collectively exhaustive. Since the union of a set of mutually exclusive and collectively exhaustive events includes all outcomes in the sample space the probability of the union of a set of mutually exclusive and collectively exhaustive events is equal to 1.

j. Intersection of events: The set of outcomes which belong to each of the events. In the die experiment if B = {1,2,3} and C = {3,4,5}, the intersection of B and C contains the outcome 3, the only outcome belonging to both B and C. The notation for the intersection of B and C would be (B and C) or $(B \cap C)$, where $\cap$ indicates intersection. $(B \cap C) = \{3\}$.

k. Independent events: When the probability of occurrence of an event is not affected in any way by the occurrence of another event. Looking again at the die toss, it is logical that the chances of rolling a 6 on a second toss is in no way influenced by the first toss since a fair die will seek an equally likely random outcome each time it is

tossed. This means that each toss result is independent of each other.

l. Dependent events: When the probability of occurrence of an event is affected by the occurrence of another event. Suppose you had a standard deck of 52 poker cards (4 suits of 13 cards each with Ace through King) and wanted to know the probability of two events:

Event 1 = {Drawing a Red Card}
Event 2 = {Drawing a Heart}

$$\mathbf{P(Event\ 1)} = \frac{\mathbf{13\ Hearts + 13\ Diamonds}}{\mathbf{52\ Cards}} = \frac{\mathbf{26}}{\mathbf{52}} = \frac{\mathbf{1}}{\mathbf{2}} = \mathbf{0.5}$$

$$\mathbf{P(Event\ 2)} = \frac{\mathbf{13\ Hearts}}{\mathbf{52\ Cards}} = \frac{\mathbf{1}}{\mathbf{4}} = \mathbf{0.25}$$

This would be an expression of each event. But if one card was drawn and you knew it was "Red" then the determination of Event 2 would change since the original population of 52 cards would be reduced to 26. Therefore the probability of Event 2 "given that" Event 1 has occurred is expressed as follows:

$$\mathbf{P(Event\ 2\ given\ that\ Event\ 1\ occurred)} = \frac{\mathbf{13\ Hearts}}{\mathbf{26\ Cards}} = \frac{\mathbf{1}}{\mathbf{2}} = \mathbf{0.5}$$

This confirms that the probability of Event 2 is affected by the occurrence of Event 1. Thus, Events 1 and 2 are dependent.

4. Types of Probabilities.

a. Simple probability: A simple probability is the probability of a single event occurring. This probability is normally determined by using one of the three probability interpretations discussed previously. For example, the probability of rolling an odd number using a single die would be determined using the classical approach of dividing the total number of outcomes (or sample points) which define the event (namely rolling a 1, 3, or 5) by the total number of outcomes (1, 2, 3, 4, 5, 6). Therefore, the probability of rolling an odd number is 3/6 or 0.5. Note that we are assuming that each outcome is equally likely (in other words, the die is fair). If we wanted to determine the probability of an officer being a West Point graduate, we would use a relative frequency approach and divide the total number of West Point graduates by the total number of officers.

b. The probability of the intersection of two or more events: This probability is the probability of the occurrence of "all" the events simultaneously. If you consider two events A and B, the probability of their intersection can be expressed as follows:

1) If A and B are independent, $\mathbf{P(A\ and\ B)} = \mathbf{P(A \cap B)} = \mathbf{P(A)} \cdot \mathbf{P(B)}$.

Let's modify the experiment of rolling a fair die to include the flip of a fair coin.

Let **A = {rolling a 4 on the single roll of a fair die} with**

$$\mathbf{P(A) = \frac{1}{6}}$$

and

B = {obtaining heads on the flip of a fair coin} with

$$\mathbf{P(B) = \frac{1}{2}}.$$

Since the outcome of the roll of the die has no effect on the outcome of the coin flip events A and B are independent and

$$\mathbf{P(A \cap B) = P(A) \cdot P(B) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12} = 0.083}$$

2) If A and B are dependent, $\mathbf{P(A \text{ and } B) = P(A \cap B) = P(A) \cdot P(B|A)}$, where P(B|A) is known as a conditional probability (more on this later).

Return to the experiment of drawing a heart and a red card on the single draw of a card from a standard deck of cards.

Event 1 = {Drawing a Red Card}
Event 2 = {Drawing a Heart}

$$\mathbf{P(Event\ 1) = \frac{1}{2}}$$

$$\mathbf{P(Event\ 2) = \frac{1}{4}}$$

If events 1 and 2 were independent, then the probability of drawing a red card and a heart would be $\frac{1}{8}$ $\left(\text{i. e.} \frac{1}{2} \cdot \frac{1}{4}\right)$, but these two events are not independent. The selecting of a red card changes the probability of Event 2. That probability is now ½.

$$\mathbf{P(Heart\ given\ that\ the\ card\ is\ red) = P(Heart|Red) = \frac{13}{26} = \frac{1}{2}}$$

and so

$$\mathbf{P(Red\ and\ Heart) = P(Red) \cdot P(Heart|Red) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}}$$

c. Conditional probabilities: A conditional probability is the probability of a specific event conditioned by the occurrence or nonoccurrence of some other event. In the above discussion of the probability of the intersection of selecting a red card and a heart on the single draw of a card from a standard deck of cards the probability of selecting a heart was conditioned by the selection of a red card.

In general the probability of A given B is expressed as **P(A|B)**, where the vertical line (|) represents the term "given." The formulas for finding a conditional probability are shown below.

$$\mathbf{P(A|B) = \frac{P(A \cap B)}{P(B)}}$$

and

$$\mathbf{P(B|A) = \frac{P(A \cap B)}{P(A)}}$$

Let's say that there is a group of seventy company grade officers who graduated from either West Point or Georgia Tech. The breakout by rank and school is shown in the table below.

	2nd Lieutenant	**1st Lieutenant**	**Captain**	**Total**
West Point	15	10	5	30
Georgia Tech	22	10	8	40
Total	37	20	13	70

From the table you can see that $\mathbf{P(GT) = \frac{40}{70} = \frac{4}{7}}$, $\mathbf{P(1LT) = \frac{20}{70} = \frac{2}{7}}$ and $\mathbf{P(GT \cap 1LT) = \frac{10}{70} = \frac{1}{7}}$. Find **P(GT|1LT).**

Using the formula $\mathbf{P(GT \mid 1LT) = \frac{P(GT \cap 1LT)}{P(1LT)} = \frac{1/7}{2/7} = \frac{1}{7} \cdot \frac{7}{2} = \frac{1}{2}}$

A conditional probability reduces the considered sample space down to that corresponding to the given event, in this case 1st Lieutenant. In calculating the conditional probability **P(GT|1LT)** you know the selected officer is a 1st Lieutenant – it's the given event. Thus, you only have to consider the set of 1st Lieutenants. Ten officers from this group graduated from Georgia Tech and so, $\mathbf{P(GT \mid 1LT) = \frac{10}{20} = \frac{1}{2}}$, which is the same value as obtained by using the formula.

Note that **P(GT|1LT)** and **P(1LT|GT)** do not have the same value.

$$\mathbf{P(1LT \mid GT) = \frac{10}{40} = \frac{1}{4}}$$

There is a relationship between conditional probabilities and independent events. If events A and B are independent, then **P(A|B) = P(A)** and **P(B|A) = P(B).**

d. The probability of the union of two or more events: This probability is the probability of one or more of the events occurring. In other words if any outcome belonging to any of the specified events occurs, then the union of those events has occurred. Considering only two events A and B, the probability of their union can be expressed as follows:

$$\mathbf{P(AorB) = P(A) + P(B) - P(AandB)} \text{ or}$$
$$\mathbf{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

For example, the probability of drawing a face card or a diamond would be

$$\mathbf{P(Face \cup Diamond) = P(Face) + P(Diamond) - P(Face \cap Diamond)}$$
$$\mathbf{P(Face \cup Diamond) = \frac{12}{52} + \frac{13}{52} - \frac{3}{52} = \frac{22}{52} = 0.42}$$

We subtract the three diamond face cards because those three cards are included in both the count of 12 face cards in the deck and the thirteen diamonds in the deck.

On the other hand, the probability of selecting a face card or an ace would be

$$\mathbf{P(Face \cup Ace) = \frac{12}{52} + \frac{4}{52} = \frac{16}{52} = 0.31}$$

The reason we only add the two individual probabilities is that selecting a face card

and selecting an ace are mutually exclusive. Thus, **P(Face and Ace) = 0.**
As was discussed earlier (see paragraph 3h above) the probability of the union of a set of mutually exclusive events is the sum of their individual probabilities.

5. Bayes' Rule.

   a. The Reverend Thomas Bayes (1702 – 1761) was an English mathematician who discovered an important relation for conditional probabilities. This relation is referred to as Bayes' Rule or Bayes' Theorem. It uses conditional probabilities to adjust calculations to accommodate new relevant information. A doctor running a test to determine the chances of a patient having a particular disease serves as an example for the use of Bayes' Rule. If the doctor knows the probability of a person having the disease and the accuracy of the test (i.e. the probability of a positive test result given that a person has the disease), he can use Bayes's Rule to determine the likelihood of his patient having the disease given a positive test result.

   b. Generally speaking, a person such as the doctor in the above example knows the probability of a person having the disease, P(A); the probability of a positive test result given that a person has the disease, P(B|A); and the probability of a false positive test result, $P(B \mid \overline{A})$, but has to find the probability of a person having the disease given a that the person had a positive test result, P(A|B). This is where Bayes' Rule comes into play. The formulas are shown below.

$$\mathbf{P(A \mid B) = \frac{P(A \cap B)}{P(B)}}$$

$$\mathbf{P(B) = P(A \cap B) + P(\overline{A} \cap B)}$$

$$\mathbf{P(A \cap B) = P(A)P(B \mid A)}$$

$$\mathbf{P(\overline{A} \cap B) = P(\overline{A})P(B \mid \overline{A})}$$

and Bayes' Rule becomes

$$\mathbf{P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(\overline{A})P(B \mid \overline{A})}}$$

c. Bayes' Rule can also be worked in tabular form.

Events	Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
A	**P(A)**	**P(B|A)**	**P(A ∩B) = P(A)P(B|A)**	**P(A|B) = P(A∩B)/P(B)**
$\overline{A}$	**P($\overline{A}$)**	**P(B|$\overline{A}$)**	**P($\overline{A}$ ∩B) = P($\overline{A}$)P(B|$\overline{A}$)**	**P($\overline{A}$|B) = P($\overline{A}$∩B)/P(B)**
			P(B) = P(A∩B) + P($\overline{A}$∩B)	

d. A Bayes' Rule Example.

Suppose it is known that 1% of the population suffers from a particular disease. A fairly simple blood test is available that has a 97% chance of providing positive results for diseased individuals, but also has a 6% chance of falsely indicating that a healthy person has the disease.

- What is the probability that a person will receive a positive blood test?
- If the blood test is positive, what is the likelihood that the individual has the disease?
- If the blood test is negative, what is the likelihood that the individual does not have the disease?

Define the Events
Let D represent the event that the person has the disease.
Let T represent the event that the test is positive.

What is given:	What you can conclude:	What is required:
• $P(D) = 0.01$	• $P(D^c) = 0.99$	• $P(T)$
• $P(T\|D) = 0.97$	• $P(T^c\|D) = 0.03$	• $P(D\|T)$
• $P(T\|D^c) = 0.06$	• $P(T^c\|D^c) = 0.94$	• $P(D^c\|T^c)$

Using the formulas:

- What is the probability that a person will receive a positive blood test?

$P(T) = P(D \text{ and } T) + P(D^c \text{ and } T)$
$= P(D)P(T|D) + P(D^c)P(T|D^c)$
$= (0.01)(0.97) + (0.99)((0.06)$
$= 0.0691$

- If the blood test is positive, what is the likelihood that the individual has the disease?

$$P(D \mid T) = \frac{P(D \text{ and } T)}{P(T)} = \frac{P(D)P(T \mid D)}{P(T)} = \frac{(0.01)(0.97)}{0.0691} = 0.1404$$

- If the blood test is negative, what is the likelihood that the individual does not have the disease?
  - First find the probability of a negative blood test.

    $P(T^c) = P(D \text{ and } T^c) + P(D^c \text{ and } T^c)$
    $= P(D)P(T^c|D) + P(D^c)P(T^c|D^c)$
    $= (0.01)(0.03) + (0.99)().94)$
    $= 0.9309$

  - Then find the probability of not having the disease given a negative blood test.

$$P(D^c \mid T^c) = \frac{P(D^c \text{ and } T^c)}{P(T^c)} = \frac{P(D^c)P(T^c \mid D^c)}{P(T^c)} = \frac{(0.99)(0.94)}{0.9309} = 0.9997$$

Using the table:

Events	**Prior Probability**	**Conditional Probability**	**Joint Probability**	**Posterior Probability**
D	P(D) = 0.01	P(T|D) = 0.97	P(T and D) = 0.0097	**P(D|T) = 0.1404**
D^c	$P(D^c) = 0.99$	$P(T|D^c) = 0.06$	$P(T \text{ and } D^c) = 0.594$	$P(D^c|T) = 0.8596$
	1.00		**P(T) = 0.0691**	1.000
D	P(D) = 0.01	$P(T^c|D) = 0.03$	$P(T^c \text{ and } D) = 0.0003$	$P(D|T^c) = 0.0003$
D^c	$P(D^c) = 0.99$	$P(T^c|D^c) = 0.94$	$P(T^c \text{ and } D^c) = 0.9306$	**$P(D^c|T^c)$ = 0.9997**
	1.00		$P(T^c) = 0.0309$	1.000

Using the table results in the same solution:

- The probability of a positive blood test is P(T) = 0.0691.
- The probability of having the disease given a positive blood test is P(D|T) = 0.1404
- The probability of not having the disease given a negative blood test is $P(D^c|T^c) = 0.997$.
-

6. Counting Formulas.

a. Many times in determining the probability of an event it is not necessary to actually identify all of the outcomes. It is only necessary to determine the number of outcomes associated with the event or to determine the total number of sample points in the sample space. Counting formulas provide a means for determining these numbers.

b. Multiplication rule: If there are *n* possible outcomes for event E_1 and *m* possible outcomes for event E_2, then there are *n x m* or *nm* possible outcomes for the series of events E_1 followed by E_2. This rule extends to outcomes created by a series of three, four, or more events. As a simple example, suppose that a person has four shirts, five pair of trousers, and three pair of shoes. How many different outfits are possible? This person can select any one of the four shirts, five pair of trousers, and three pair of shoes. There are 4 x 5 x 3 or 60 different possible outfits.

c. Finding the number of ordered arrangements of *n* items: Sometimes it is necessary to determine the total number of different ordered arrangements possible when arranging all of a group of *n* items.

This can best be demonstrated through an example. Let's say that five people must be seated in a row. The multiplication rule can be applied. There are five possibilities for the filling the first seat. Now that a selection has been made, there are four possibilities for filling the second seat. Having filled the first two seats there are now three possibilities for filling the third seat. Continuing in this manner and using the multiplication rule there are a total of 5 x 4 x 3 x 2 x 1 or 120 possible seating arrangements.

This multiplication pattern is an example of the multiplication indicated by the factorial notation *n!* or in this example 5!.

! is read "factorial"
5! is read "5 factorial"
5! = 5 x 4 x 3 x 2 x 1 = 120

The factorial notation can be summarized as follows:
$n! = n(n-1)(n-2)\ldots 1$
0! = 1 by special definition
1! = 1

In summary, if *all* of the members of a group of *n* items must be arranged, then there are *n!* possible arrangements.

d. Permutations: The number of ways to *arrange in order n* distinct objects, taking them *r* at a time, is given by the formula shown below.

$$ {}_{n}P_{r} = \frac{n!}{(n-r)!} $$

The notation reads "the permutation of n items taken r at a time."

In expanding the seating example from above, assume there are now eight people, but only five chairs. How many ways can five of the eight people seat themselves in the five chairs? In this case the order of the arrangement is important. The number of arrangements can be found as follows:

$$ {}_{8}P_{5} = \frac{8!}{(8-5)!} = \frac{8!}{3!} = \frac{40{,}320}{6} = 6{,}720 $$

There are 6,720 ordered seating arrangements for seating five out of eight people.

e. Combinations: If the order of the arrangement is not important, then the combination rule can be used to determine the total number of possibilities. The combination rule is given by the formula shown below.

$$ {}_{n}C_{r} = \binom{n}{r} = \frac{n!}{(n-r)!r!} $$

The notation ${}_{n}C_{r}$ and $\binom{n}{r}$ both mean to take a combination of *n* items *r* at a time.

Let's look at an example. A committee of five people must be selected from a group of eight people. How many different committees are possible? In this case the order of selection is irrelevant. It really does not matter in which order these five people are selected. So the combination rule is appropriate and the number of possible committees can be found as follows:

$$ {}_{8}C_{5} = \binom{8}{5} = \frac{8!}{(8-5)!5!} = \frac{8!}{3!5!} = \frac{40{,}320}{6(120)} = \frac{40{,}320}{720} = 56 $$

Notice the difference between the concepts of permutations and combinations. Permutations consider both groupings and order. Order counts. Combinations consider only groupings. The order of the arrangement is not important.

7. Probability Distributions.

a. Random variable: A random variable is defined as the numerical outcome of a random experiment.

b. Probability distribution: A probability distribution is an assignment of probabilities to the specific values of a random variable or to a range of values of a random variable. Suppose a random experiment consist of flipping a fair coin three times and the random variable is defined as the number of heads obtained on these three flips. There are eight possible outcomes for this experiment.

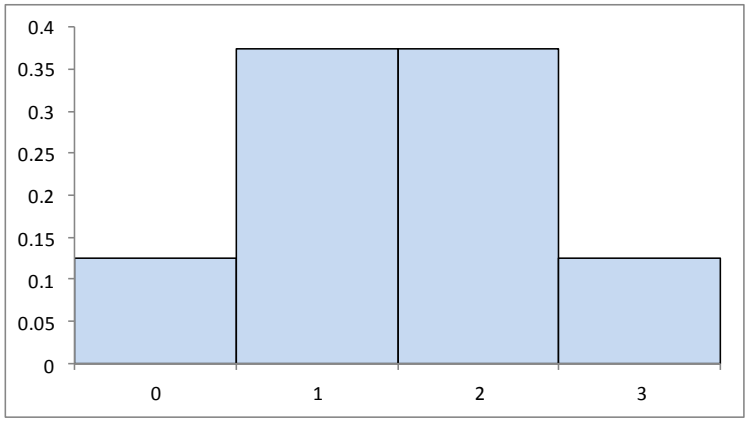
T-T-T T-H-T T-T-H H-T-T
H-T-H H-H-T T-H-H H-H-H

The possible values for the random variable and their associated probabilities are shown in the table below.

X (Number of Heads)	**P(X)**
0	1/8 = 0.125
1	3/8 = 0.375
2	3/8 = 0.375
3	1/8 = 0.125
	$\sum P(X) = 1.000$

This table represents the probability distribution for the random variable X, the number of heads obtained on the flipping a fair coin three times. Note that the probability distribution accounts for all the possible values of X and the probability of occurrence for each of those values. Also note that the sum of the probabilities is equal to 1.

This probability distribution can also be represented graphically.



This graph is simply a relative frequency histogram. The horizontal axis displays all of the possible outcomes (i.e. 0, 1, 2, 3), while the vertical axis displays the

probability of occurrence. The heights of the bar reflect the probability of occurrence for each of the outcomes. Note that the total area under the curve is equal to 1.

c. Types of probability distributions.

1) Discrete: A discrete probability distribution assigns probability values to a set of finite or infinitely countable integers usually, but not necessarily, beginning with zero. Think of discrete as things that can be counted. Consider the following examples:

a) The above example of tossing a coin three times and defining the random variable, X, to be the number of heads is an example of a discrete probability distribution. This distribution serves as an example of a discrete probability distribution with a countable finite set of events.

b) A second example involves firing a missile at a target and defining the random variable X to be the trail number on which the target is first hit. In this case X can take on an infinitely countable set of values (i.e. 1^{st} trial, 2^{nd} trial, 3^{rd} trial, …, 180^{th} trial, …). The probability distribution would also include the probability associated with each of these infinitely countable set of values.

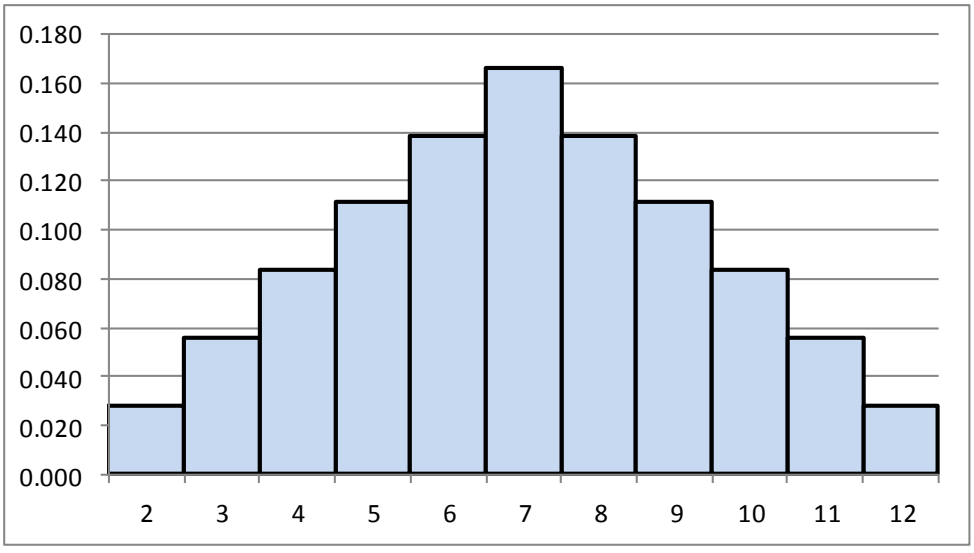
2) Continuous: A continuous random variable X takes all values in an interval of numbers. Think of a continuous random variable as something that is measured such as time, height, or weight. The probability distribution is described by a density curve. With a continuous random variable it is impossible to find the probability of the variable being exactly equal to some value. For example, it is impossible to define the probability of a person being exactly 6 feet tall. Only intervals of values have positive probability and this probability is equal to the area under the density curve corresponding to the interval of values. For example, the probability of a person being between 6 feet and 6 ½ feet tall would correspond to the area under the density curve between 6 and 6 ½ feet.

d. Cumulative Probability Distributions. A cumulative probability distribution is a distribution of probabilities assigned to an accumulation of events. In effect it provides the probability of a variable being less than or equal to some specified value. The cumulative probability distribution can be either discrete or continuous.

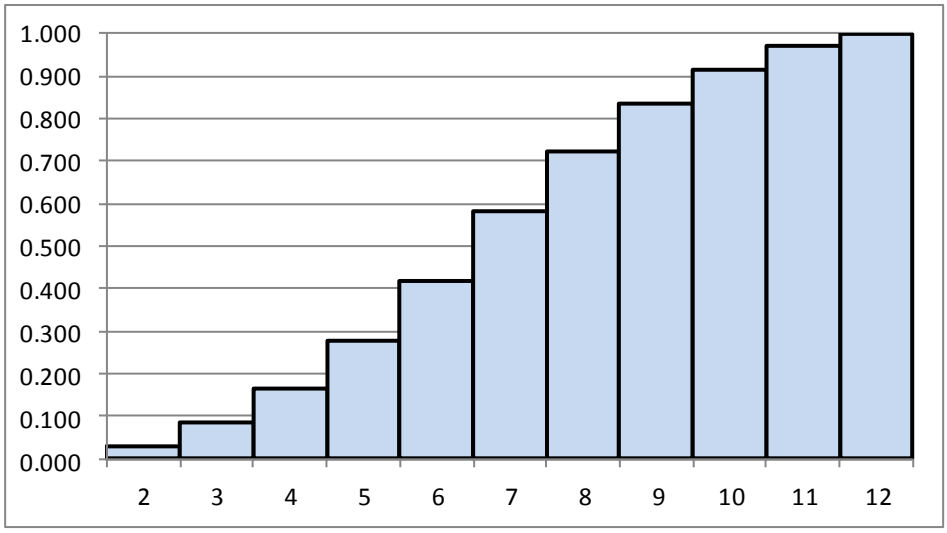
1) Consider the example of throwing two dice and recording the sum of the two dice. The probabilities associated with each outcome are given in the left-hand table below. The cumulative probability distribution is shown in the right-hand table.

Outcome	**Probability**		**Outcome**	**Cumulative Probability**
2	1/36 = .028		2 or less	1/36 = .028
3	2/36 = .056		3 or less	3/36 = .083
4	3/36 = .083		4 or less	6/36 = .167
5	4/36 = .111		5 or less	10/36 = .278
6	5/36 = .139		6 or less	15/36 = .417
7	6/36 = .167		7 or less	21/36 = .583
8	5/36 = .139		8 or less	26/36 = .722
9	4/36 = .111		9 or less	30/36 = .833
10	3/36 = .083		10 or less	33/36 = .917
11	2/36 = .056		11 or less	35/36 = .972
12	1/36 = .028		12 or less	36/36 = 1
$\sum P(X) =$	36/36 = 1			

2) Each of the discrete outcomes has one discrete probability value associated with it. These probabilities are plotted against their respective outcomes in either block or spike form to give the discrete probability distribution shown below.



3) Now the probabilities associated with the accumulated outcomes are shown in the right-hand table above. The resulting cumulative probability distribution is shown in the graph below.



The cumulative probability distribution can be used to find the probability of the sum being less than or equal to some specified value. For example, the probability of the sum being 6 or less is 0.417, which can be read directly from the table or the graph.

To find the probability of obtaining a sum greater than 6 subtract the probability of getting a sum of 6 or less from 1. The probability of a sum greater than 6 would be 0.583 (i.e. 1 – 0.417).

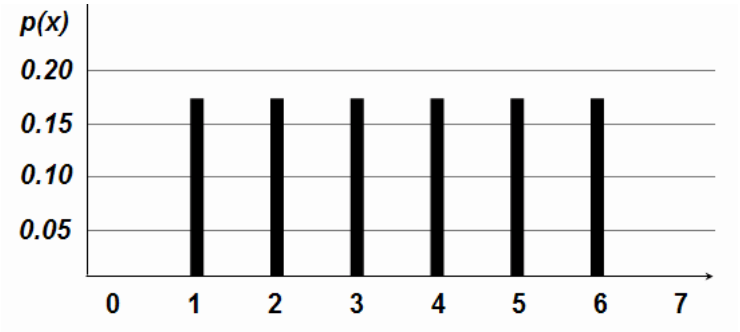
e. Many phenomena and common everyday occurrences will tend to have a certain type of probability distribution associated with them. For example, in situations where individuals are waiting for some type of service (e.g. the bank, a fast food restaurant, etc.), the arrival rate of individuals joining the line tends to follow a Poisson distribution. There are various types of probability distributions. The remainder of this chapter will cover examples of some of the more commonly used discrete and continuous probability distributions. Most statistics books will have additional information concerning other probability distributions.

8. Discrete Probability Distributions.

a. Discrete Uniform Distribution

1) Each event in the discrete uniform distribution has the same chance of occurrence or probability.

2) As an example, roll a fair die. There are six possible outcomes, rolling a 1, 2, 3, 4, 5, or 6. Each outcome has the same chance of occurring, that is a probability of 1/6. The graph would therefore look like:



3) The mean of the discrete uniform distribution is $\mu = \frac{a+b}{2}$, where *a* and b are the minimum and maximum values, respectively. The variance is

$$\sigma^2 = \frac{(b-a)^2 - 1}{12}.$$

4) For any value of x $P(x) = \frac{1}{n}$, where *n* is the number of outcomes associated with the random variable. In the above example, there are six possible outcomes associated with the single roll of a fair die. Therefore, the probability of any outcome is $\frac{1}{n} = \frac{1}{6} = 0.167$.

5) An Excel template entitled "Distributions Template" serves as an example of using Excel to work with probability distributions (see Appendix B for instructions on how to obtain access to the template). For the discrete uniform distribution the required inputs (shown in yellow in the template) are the minimum value, *a*, the maximum value, *b*, and the value of x. The template will return the following probabilities:

$P(X = x)$
$P(X \geq x)$ and $P(X < x)$
$P(X > x)$ and $P(X \leq x)$

For example, finding the probabilities associated with rolling a three on a single roll of a fair die require the inputs of $a = 1$, $b = 6$, and $x = 3$ as shown in the figure below, taken from the Excel template. The template returns the values shown in the figure.

a	1		
b	6		
x	3	P(X = x)	0.1667
		P(X ≥ x)	0.6667
		P(X < x)	0.3333
		P(X > x)	0.5000
		P(X ≤ x)	0.5000

b. Binomial Distribution.

1) The binomial probability distribution is used to find the probability associated with X number of discrete successes occurring while conducting *n* number of trials of a certain experiment when the probability of success is constant for each trial. For instance, if the probability of hitting an enemy target is 0.85 with a certain weapon system, then the binomial distribution can be used to find the probability of hitting the target 7 out of 10 times.

2) The following conditions must be in place in order to conduct a binomial experiment and compute probabilities.

   a) There are *n* independent trials, in this case n = 10.

   b) There are only two outcomes per trial, a success or a failure. In this case success is hitting the enemy target, and failure is not hitting the enemy target.

   c) The probability of success is constant for all trials and is designated by *p*. In this case the probability of success corresponds to the probability of hitting an enemy target, with $p = 0.85$. The probability of failure is designated by *q*, with $q = 1 - p = 1 - 0.85 = 0.15$.

   d) X is the number of successful outcomes when the experiment is completed. In this case, X = 7 successful outcomes or hits on the enemy targets.

3) The binomial formula for finding P(X = x) is $P(X = x) = {}_nC_x p^x q^{n-x}$. In this example n = 10, p = 0.85, q = 0.15, and x = 7.

$$P(X = 7) = {}_{10}C_7(0.85)^7(0.15)^3 = 120(0.32058)(0.00375) = 0.1298.$$

4) When using the Excel template the inputs (shown in yellow) are *n*, *p*, and *x*. The outputs are the same as for the discrete uniform distribution and are displayed in the figure.

n	10		
p	0.85		
x	7	P(X = x)	0.1298
		P(X ≥ x)	0.9500
		P(X < x)	0.0500
		P(X > x)	0.8202
		P(X ≤ x)	0.1798

5) The mean for the binomial distribution is $\mu = np$ and the variance is $\sigma^2 = npq$. For this example, $\mu = np = 10(0.85) = 8.5$ and $\sigma^2 = npq = 10(0.85)(0.15) = 1.275$.

c. Poisson Distribution.

1) The Poisson distribution is used to find the probability of the number of outcomes occurring during a given time interval or in a specified region when the average number of occurrences is known.

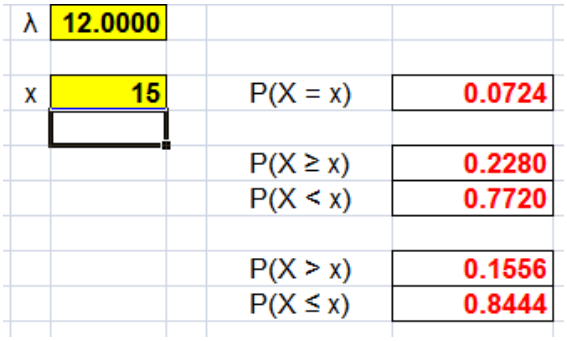
2) The Poisson distribution formula is $P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$, where λ is the average number of occurrences within a given time interval or region.

3) For example, given that the average number of customers entering a bank is two every ten minutes find the probability that 15 will enter in an hour.

First, convert the average number of occurrences from 2 per every 10 minutes to 12 per hour.

Then $P(X = 7) = \frac{e^{-12}(12)^{7}}{7!} = 0.0724$

4) When using the Excel template the inputs (shown in yellow) are λ and x. Once again the outputs are the same as for the discrete uniform distribution.



5) The mean and variance for the Poisson distribution are both λ.

9. Continuous Probability Distributions.

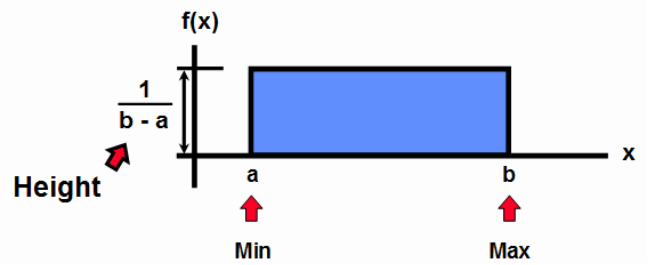
a. Continuous Uniform Distribution.

1) The continuous uniform distribution is a rectangular distribution where the probability of falling in an interval of fixed length [a, b] is constant.

2) The density function of the continuous uniform random variable X on the interval [a, b] is

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{elsewhere} \end{cases}$$

3) This density function forms a rectangle with base $b - a$ and height $\frac{1}{b-a}$ (See figure below.)



4) Note that the total area under the curve is equal to 1 and the probability of the random variable X taking on a value within any interval $[x_1, x_2]$ is equal to the area under the curve between x_1 and x_2. This probability can be found using the formula $P(x_1 \le x \le x_2) = \frac{x_2 - x_1}{b - a}$.

5) For example, say that the time to calibrate a GPS is uniformly distributed between 1 and 4 minutes. Find the probability that the calibration will take between 3 and 3.5 minutes.

$$P(3 \le x \le 3.5) = \frac{3.5 - 3}{4 - 1} = \frac{1}{6} = 0.167$$

The Excel "Distributions Template" includes a set of continuous probability distributions. The inputs are *a*, *b*, and *x*. The outputs are *P(X < x)* and *P(X> x)*. Remember that with continuous probability distributions P(X = x) = 0. The inputs *Lo* and *Hi* are required to solve a problem like this one. The template returns the output for *P(Lo ≤ X ≤ Hi)*. The Excel output is displayed in the figure.

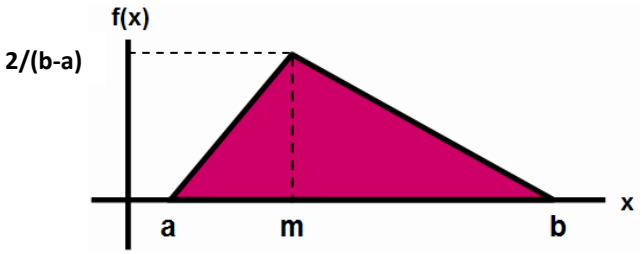
a	1		
b	4		
x	3	P(X < x)	0.666667
		P(X > x)	0.333333
Lo	3	P(Lo ≤ X ≤ Hi)	0.166667
Hi	3.5		

b. Triangular Distribution.

1) The triangular distribution is shaped like a triangle with a lower bound, *a*, and upper bound, *b*, and a mode, *m*. It, like the continuous uniform distribution, is often used to model subjective estimates such as a cost estimate where a subject matter expert provides estimates of the minimum, maximum, and most likely costs.

2) The density function for the triangular distribution is

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(m-a)} & a \le m \le b \\ \dfrac{2(b-x)}{(b-a)(b-m)} & m < x \le b \\ 0 & \text{otherwise} \end{cases}$$



The cumulative density function can be used to calculate $P(X \le x)$.

$$\text{cdf}: \begin{cases} \dfrac{(x-a)^2}{(b-a)(m-a)} & a \le x \le m \\ 1 - \dfrac{(b-x)^2}{(b-a)(b-m)} & m \le x \le b \end{cases}$$

3) For example, an engine can usually be overhauled in 2 hours. However, it can take as little as 1 hour and as long as 4 hours. What is the probability that it will be completed between 1.5 and 3 hours? In other words, find $P(1.5 \le x \le 3)$. Given: a = 1; b = 4; m = 2

$P(1.5 \le x \le 3) = P(x \le 3) - P(x \le 1.5)$

First, solve $P(x \le 3) = 1 - \dfrac{(4-3)^2}{(4-1)(4-2)} = 0.8333$

Then solve $P(x \le 1.5) = \dfrac{(1.5-1)^2}{(4-1)(2-1)} = 0.0833$

Finally, $P(1.5 \le x \le 3) = P(x \le 3) - P(x \le 1.5) = 0.8333 - 0.0833 = 0.75$

4) The above can be solved more easily using the Excel "Distributions Template." The inputs are the minimum value, *a*, the maximum value, *b*, and the mode, *m*. The outputs are *P(X < x)* and *P(X> x)*. The inputs *Lo* and *Hi* are required to solve this problem. The template returns the output for *P(Lo ≤ X ≤ Hi)*. The Excel output is displayed in the figure.

a	1		
Mode	2		
b	4		
x	3	P(X < x)	0.833333
		P(X > x)	0.166667
Lo	1.5	P(Lo ≤ X ≤ Hi)	0.750000
Hi	3		

5) The mean and variance for the triangular distribution are

$$\mu = \frac{a + m + b}{3}$$

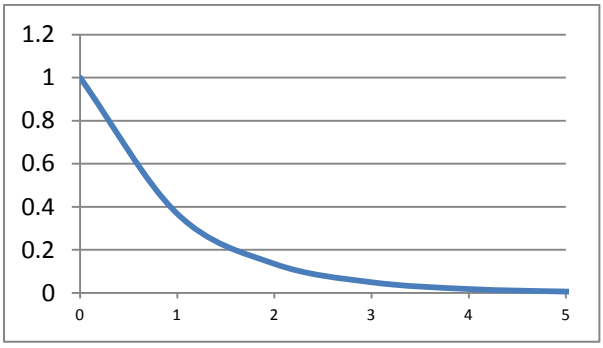
$$\sigma^2 = \frac{a^2 + m^2 + b^2 - ab - am - mb}{18}$$

c. Exponential Distribution.

1) An exponential distribution is often used to represent the distribution of the time that elapses before the occurrence of some event. For example, this distribution has been used to represent the period of time for which a machine will operate before breaking down, the period of time required to service a customer, or the period between the arrivals of two customers.

2) If the events being considered occur in accordance with a Poisson process, then the waiting time until an event occurs and the time between any two successive events will follow the exponential distribution.

3) The probability density function is $f(x) = \lambda e^{-\lambda x}$. The graph of an exponential density function with $\lambda = 1$ is shown below.



4) The cumulative density function $1 - e^{-\lambda x}$ can be used to calculate $P(X \leq x)$.

5) The example described in the section on the Poisson distribution set the average number of customers entering a bank at 2 every 10 minutes or 0.2 customers every minute. Find the probability that the next customer will arrive within 2 minutes? Since the arrivals follow the Poisson process, the time between arrivals or until the next arrival follows the exponential distribution. If the average number of arrivals per minute is 0.2 customers, then the expected time between arrivals is 1/0.2 or 5 minutes. For this problem $\lambda = 0.2$. The cumulative density function to find $P(X \leq 2)$.

$$P(X \leq 2) = 1 - e^{-0.2(2)} = 1 - e^{-0.4} = 0.3297$$

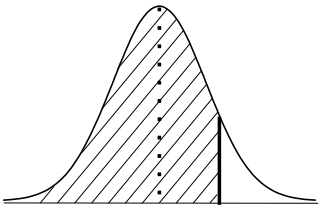
6) The Excel "Distributions Template" can be used to solve this problem. The inputs are λ and x with the outputs being P(X < x) and P(X > x). The Excel output is shown in the figure.

λ	0.2000		
x	2.0000	P(X < x)	0.329680
		P(X > x)	0.670320

7) The mean and variance for the exponential distribution are $\mu = \frac{1}{\lambda}$ and $\sigma^2 = \frac{1}{\lambda^2}$.

d. Normal Distribution.

1) Repeated, independent measurements of many physical properties such as weights, heights, age, cost, etc. tend to follow a normal distribution. All normal distributions have the same basic bell curved shape defined by two parameters, the distribution mean, μ, and the distribution standard deviation, σ. The mean identifies its location along the horizontal axis, while the standard deviation defines its shape by controlling the spread of the curve. The larger the standard deviation the more spread out the curve. A normal curve is displayed in the figure.



2) The shaded area under the curve represents the probability of a cumulative event. The total area under the curve is equal to one. Since the distribution is symmetric about the mean, 50% of the area is above the mean or to the right of the mean and 50% of the area is below or to the left of the mean. The area under the curve between any two values [a, b] is equivalent to the probability of the random variable taking on a value between a and b.

3) There are many different normal random variables with different values for their means and standard deviations. Each of these random variables would have a different bell-shaped curve determined by their respective mean and standard deviation, but all of these distributions have one thing in common and that is that all normal distributions are the same if measured in units of standard deviations of size σ about the mean μ as center. For example, the percentage of the area under the curve between the mean and two standard deviations above the mean is the same for all normal distributions. This commonality can be used to simplify solving probability problems involving the normal distribution. This is accomplished by converting the value of the normally distributed random variable X to its equivalent standard normal value known as a Z score.

4) The standard normal or Z distribution has a mean μ equal to 0 and a standard deviation σ equal to 1. The Z formula shown below is used to convert a value, X = x, to its equivalent Z score.

$Z = \frac{x - \mu}{\sigma}$, where x = the raw score from the normal distribution

μ = the mean of the distribution

σ = the standard deviation of the distribution

The Z score is a distance score which represents the distance that any given raw score *x* lies from the mean in terms of standard deviation units. For example, if a raw score of X = 10 converts to a Z score of 1.25, this implies that the score 10 is 1.25 standard deviations away from the mean. A Z score of 0 indicates that the raw score value it represents is the same as the mean. A positive Z score implies the raw score is Z standard deviations above the mean, while a negative Z score implies the raw score is Z standard deviations below the mean.

5) Tables have been developed to provide areas under the standard normal curve. The cumulative standard normal table at the back of the reference book (see Appendix C) provides cumulative areas for Z scores ranging in value from -3.49 to +3.49. The cumulative area is the area below or to the left of a Z score. To use this standard normal table first convert the raw score *x* to its equivalent Z score. Then enter that table with that Z score to find the area to the left of the Z score. This area is equivalent to the probability of Z having a value less than the calculated Z score. This, in turn, is equivalent to random variable X having a value less than *x*.

   a) In using the standard normal table (see the figure below) the Z score consists of three digits, an integer digit and two decimal place digits (e.g. Z = -2.95).

   b) The first two digits (i.e. the integer digit and the tenths place digit identify the row, while the third digit (i.e. the hundredths place digit) identifies the column.

   c) The intersection of the row and column provide the area to the left of the given Z score or the probability of Z being less than the given Z score.

   d) For example, P(Z < -2.95) = 0.0016.

z	0.00	0.01	0.02	0.03	0.04	0.05
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040

6) To see how this works consider the following example.

a) The heights of a large group of students are normally distributed with a mean of 72 inches and a standard deviation of 6 inches. The notation representing this distribution is X ~ n(72,6) and reads "X is normally distributed with a mean of 72 and a standard deviation of 6."

b) In randomly selecting one student from this group find the probability that the student is less than 66 inches tall. That is, find P(X < 66).
(1) First, convert X = 66 to its corresponding Z score.

$$Z = \frac{x - \mu}{\sigma} = \frac{66 - 72}{6} = \frac{-6}{6} = -1$$

A Z score of -1 implies that 66 inches is one standard deviation below the mean of 72 inches.

(2) Second, use the cumulative standard normal table (see Appendix C) to find P(Z < -1.00).

(a) Move down the Z column of the table to the row labeled -1.0.

(b) Then move across the columns to the column labeled 0.00.

(c) Then read the value displayed at the intersection of row -1.0 and column 0.00 to obtain the value 0.1587.

z	0.00	0.01	0.02	0.03
-3.4	0.0003	0.0003	0.0003	0.0003
-3.3	0.0005	0.0005	0.0005	0.0004
-3.2	0.0007	0.0007	0.0006	0.0006
-3.1	0.0010	0.0009	0.0009	0.0009
-3.0	0.0013	0.0013	0.0013	0.0012
-1.4	0.0808	0.0793	0.0778	0.0764
-1.3	0.0968	0.0951	0.0934	0.0918
-1.2	0.1151	0.1131	0.1112	0.1093
-1.1	0.1357	0.1335	0.1314	0.1292
-1.0	0.1587	0.1562	0.1539	0.1515

(d) Therefore, P(Z < -1.00) = 0.1587 and P(X < 66) = 0.1587.

(e) Thus, the probability of randomly selecting a student with a height less than 66 inches is 0.1587 or 15.87%.

7) Now find the probability of selecting a student at random whose height is more than 66 inches. That is find P(X > 66).

   a) P(X > 66) = 1 – P(X < 66) = 1 – 0.1587 = 0.8413.

   b) Therefore, the probability of selecting a student at random whose height is more than 66 inches is 0.8413 or 84.13%.

8) Find the probability of selecting a student at random whose height is between 66 and 80 inches.

   a) The probability of a student being shorter than 80 inches, P(X < 80), takes into account a student ranging in height from 0 inches up through 80 inches tall.

   b) The probability of a student being shorter than 66 inches, P(X < 66), includes a student ranging in height from 0 inches through 66 inches tall.

   c) Therefore, the probability of a student having a height between 66 and 88 inches is simply the difference between these two probabilities.

   d) The problem can be rewritten as P(66 < X < 80) = P(X < 80) – P(X < 66).

   e) To solve this problem first convert each of these heights to their respective Z scores.

$$Z_{80} = \frac{80-72}{6} = \frac{8}{6} = 1.33$$
$$Z_{66} = \frac{66-72}{6} = \frac{-6}{6} = -1.00$$

   f) Finding these Z scores converts the problem from

$$P(66 < X < 80) = P(X < 80) - P(X < 66)$$
to
$$P(-1.00 < Z < 1.33) = P(Z < 1.33) - P(Z < -1.00)$$

   g) Using the standard normal table find P(Z < 1.33) and P(Z < -1.00).

      (1) It is known from the last problem that P(Z < -1.00) = 0.1587.

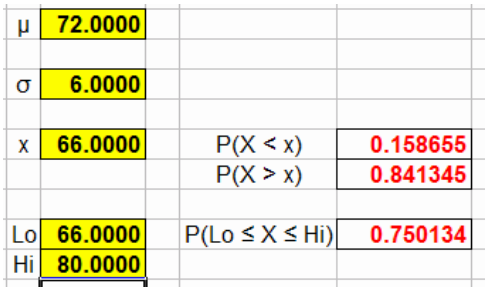
      (2) P(Z < 1.33) = 0.9082. See the figure below.

z	0.00	0.01	0.02	0.03	0.04	0.05
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265

h) Then P(-1.00 < Z < 1.33) is the difference between these two areas under the curve or P(-1.00 < Z < 1.33) = 0.9082 – 0.1587 = 0.7495.

i) Thus, the probability of randomly selecting a student with a height between 66 and 80 inches tall is 0.7495 or 74.95%.

9) These problems can be solved using the Excel "Distributions Template" as shown in the figure below. The inputs are μ, σ, and x along with the Lo and Hi values.



The successful completion of the above three problems relies on the premise that the cumulative table values always provide the reader with the cumulative probability or the probability that the random variable X is less than some given value. To find the probability that X is greater than some value, one must subtract the cumulative percentage or the probability that X is less that this given value from 1. To find the probability that X takes on a value between two specified values subtract the smaller cumulative percentage from the larger cumulative percentage.

10. Summary.

    a. Probability theory offers a way to quantitatively express one's point of view regarding the occurrence of random events. It will not establish what will happen, only the possibility that something will happen with some theoretical frequency. Since there are many laws and types of distributions from which to select, this point of view is ultimately subjective but helpful in providing useful information, which may contribute to decision-making.

    b. Real analyses of explaining or predicting the behavior of random phenomena begin with reasonable assumptions based either on theory or historical data. Probability theory and methods are simply tools, which may be used to create a working model of that random phenomenon. Probability methods will provide answers to questions, however. The results only provide insights as to what the outcomes might be if the assumptions are stable and reasonable. The actual decision-making is left to the decision maker.

11. For more information see: Walpole, Ronald E., et al. *Probability and Statistics for Engineers and Scientists,* 8th ed. Upper Saddle River, NJ: Prentice Hall, 2007.

SECTION FIVE
INFERENTIAL STATISTICS
(Return to Table of Contents)

1. Introduction.

As discussed in the descriptive statistics section, Inferential Statistics is the branch of statistics concerned with using sample data to make an inference about a population. Though we typically collect sample data, our interest is to analyze the entire population. This section will discuss several methods to do such.

2. Central Limit Theorem. The central limit theorem has two parts to it:

a. If **x** is a random variable with a normal distribution whose mean is **μ** and standard deviation is **σ**, and $\bar{x}$ is the sample mean of random samples of size **n** taken from the distribution then :

- The $\bar{x}$ distribution is a normal distribution.
- The mean of the $\bar{x}$ distribution is **μ**.
- The standard deviation of the $\bar{x}$ distribution is $\frac{\sigma}{\sqrt{n}}$.

Therefore, if we can deduce that the x distribution is normal, then we can take small samples for the sampling distribution and it will be normal.

b. If **x** is a random variable with any distribution whose mean is **μ** and standard deviation is **σ**, and $\bar{x}$ is the sample mean of random samples of size n taken from the distribution then:

- The $\bar{x}$ distribution approaches a normal distribution as n increases without limit (statistical studies have shown that large enough sample sizes are at least 30).
- The mean of the $\bar{x}$ distribution is **μ**.
- The standard deviation of the $\bar{x}$ distribution is $\frac{\sigma}{\sqrt{n}}$.

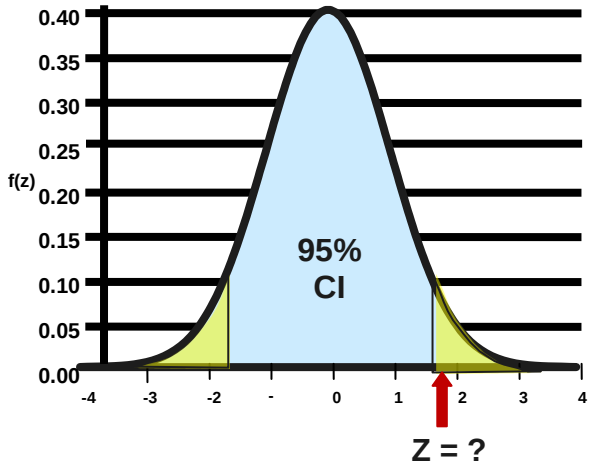
Therefore if you can take large samples to create the sampling distribution, you do not have to worry what the population distribution is as the sampling distribution will be approximately normal.

3. Confidence Intervals. Typically, we cannot afford to collect many samples and take one sample to make inferences about the population. Because the sample mean, $\bar{x}$, is not a good estimator of the population mean, we create a range of values in which we believe, with a certain degree of confidence, that the population mean will fall. This range of values is called a confidence interval and it has a **1 - α** level of confidence, where **α** is the amount of acceptable risk that the mean is not in the interval.

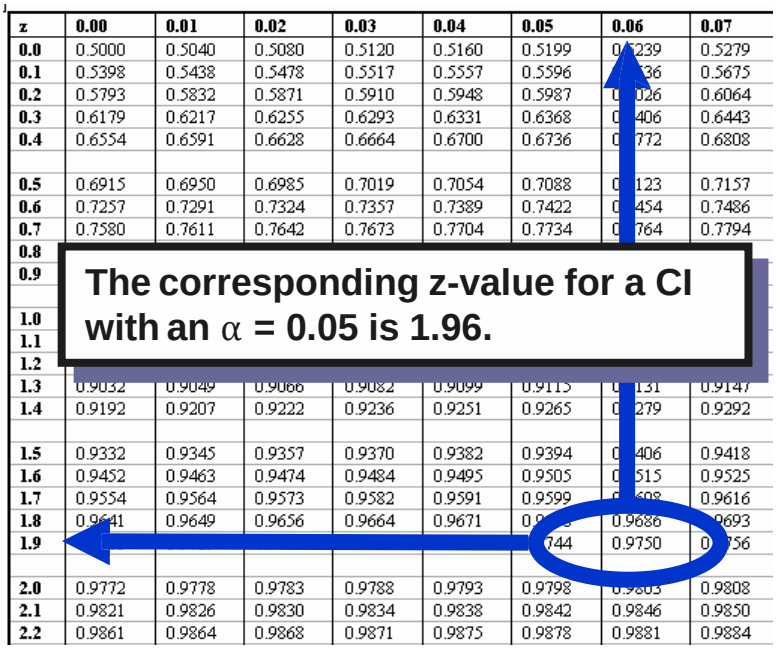
a. The formula for a confidence interval is given by $x - E < \mu < x + E$, where **E** is the amount of error from the sample mean and is given by the formula $E = z_{\alpha/2} \frac{s}{\sqrt{n}}$. You may not realize that you have seen confidence intervals many times. For example, national polls might indicate that 45% of the population approves of a presidential candidate's policy. However, another number is also given which is called an error term - say $\pm$ 3%. What this truly means is that a random sample was taken and they believe, with 95% confidence, that the true acceptance level is between 42% to 48%. In this case all the numbers are in percentages since we are discussing proportions. The error term will be in the same units as the sample mean.

b. So what is **Z** and where does it come from? Every normal distribution has a mean (μ) and a standard deviation (σ) associated with it. To avoid the use of calculus for many calculations, a table was created to find probabilities of normal distributions. A standard normal distribution was created which has a mean of **0** and a standard deviation of **1**. The values taken are **no longer x values**, **but z values**, sometimes called **z scores**. A z score is then interpreted as the number of standard deviation you are from the mean, a positive score to the right and negative score to the left.

c. With the standard normal table with can search for probabilities of z scores or vice versa. To create a confidence interval, a **z** score, associated with the confidence you desire to have, must be found. As an example, let's walk through finding a score for a 95% confidence interval. First, if we draw the distribution it will look like this:



Note that we want to create an interval where 95% of the data is in the middle of the distribution. Therefore 5% is in the tails or 2.5% in each tail. We want to search for a **z** value that is at the right end of the interval desired. We will be using a cumulative table; that is the probabilities are from - ∞ to the desired **z** score. Therefore, to find the z score associated with a 95% interval, we have to add the 2.5% on the left tail. So, our search will actually be for **97.5%** in the table.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794
0.8								
0.9								
1.0								
1.1								
1.2								
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693
1.9						0.9744	0.9750	0.9756
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884

The corresponding z-value for a CI with an α = 0.05 is 1.96.

An extract of a **z** table would look like this: (see App. C.)

Therefore, we will use a **z** score of 1.96 to create a 95% confidence interval.

d. To continue, let's take a look at an example.

For warranties to remain current, average mileage of vehicle fleet cannot exceed 12,000 miles per year. A **random sample (n) of 100** vehicles, at an Army post, shows that the vehicles are **driven an average of 12,500 miles** per year with a **sample standard deviation (s)** of **1,735** miles. Assuming a normal distribution, construct a 95% confidence interval for the average number of miles a vehicle is driven.

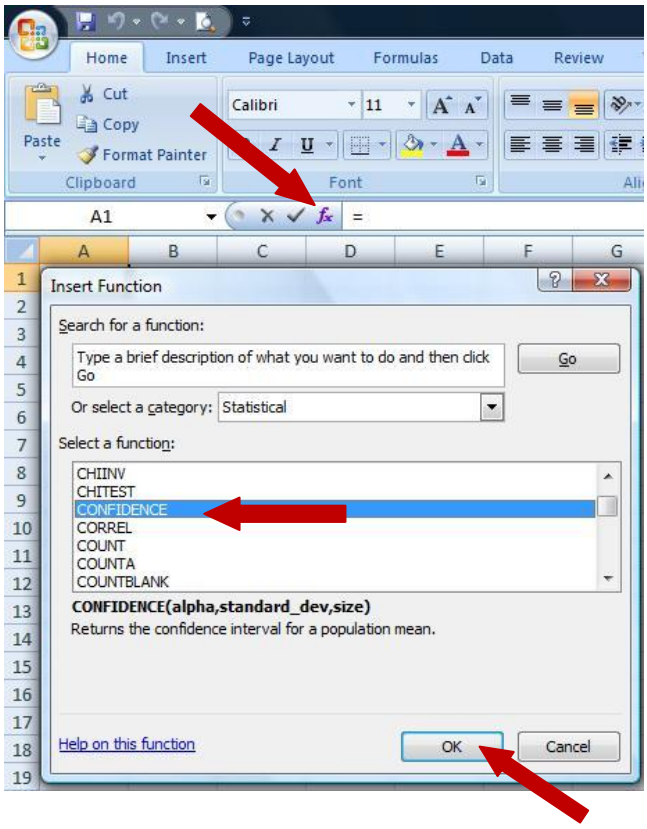
First we calculate our error: $\boldsymbol{E = Z\ \frac{s}{\sqrt{n}} = 1.96 * \frac{1735}{\sqrt{100}} = 340.05}$

Using an error of 340 miles we can calculate the interval:

$CI_{95\%} = 12{,}500 \pm 340$ or $CI_{95\%} = 12{,}160 < \mu < 12{,}840$

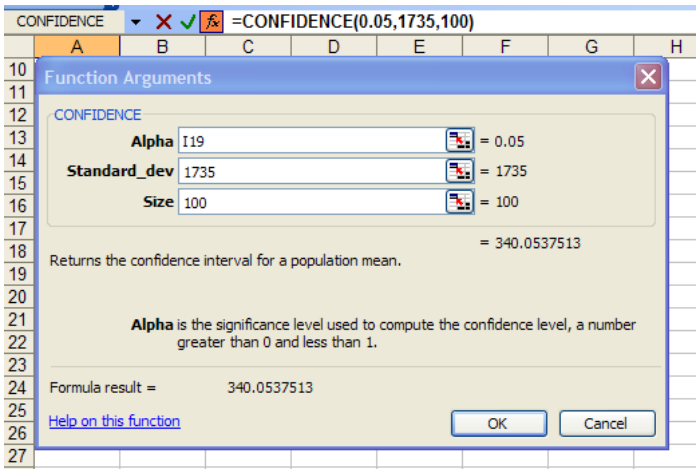
and it is said: **I am 95% confident that the average distance vehicles are driven at the base is between 12,260 and 12,840 miles/year**.

e. Let's now look at this example in Excel 2007®. Open Excel and look for the **f_x button** in the menu bar (see arrows in image on right) and click on it; this will bring up a box with all available functions (in older versions of Excel® you may have the f_x button in a different location than appears on the right and the drop down box will look slightly different). Select the **category - Statistical** and then search for the function **CONFIDENCE**. Click on **CONFIDENCE** to highlight the function and then click-on the **OK** button. The **Function Arguments** menu will appear asking for information.



The first box asks for **Alpha (α)** (note: it does not ask for the confidence level), the **Alpha is 1-Confidence Level**, (write this in decimal form not percent). For a desired confidence of 95%, the **alpha is** 5% or **0.05**. Next, enter the **standard deviation** (**1735**) and **size of your sample** (**100**). When you have entered the data click-on the **OK** button.

You should **see** the **result** in **Cell A1** – which Excel defaults to when you open your spreadsheet.



Also note that the result is not a confidence interval but the error for the interval. To create the interval, simply subtract and add the error to the sample mean as we did in d. above.

f. Another way to do this in Excel 2007® is to create a template with the formulas embedded. Each time you have to perform these operations the template is available for use.

We created a template to work with this section that will compute confidence intervals and hypothesis tests. See appendix B for the location to download the template. To use this template for confidence intervals, you do not need to fill in all the information. Only the Sample mean, Standard deviation, Sample size and Alpha level are required (note this is the same information required above except that the sample mean is required to create the actual interval). Once the information is filled in, the confidence interval is given. One advantage to using this is that you can also cell reference your inputs rather than type them in. For example, if you have a set of data, you can use the Data Analysis tool as we did in descriptive statistics to find the mean and standard deviation and then cell reference those cells into the template. If you recall in the Data Analysis tool, there is an option for Confidence Intervals, if you check that option and enter the confidence you desire, you will get within the output a cell with the error for your interval. However, you need to be aware that the Data Analysis tool uses the T Distribution to compute the error and not the normal distribution. The T Distribution is typically used for small sample sizes. I would not be too concerned since the T Distribution approaches the normal as n gets bigger. The template we created also allows the creation of confidence intervals using the T Distribution.

2	**Normal Distribution**			
3				
4	**Type Tail Test 1=Two, 2=Left, 3=Right**			
5	**Hypothesized Mean**		**Zcritical**	1.959964
6	**Sample mean**	12500	**T.S.**	
7	**Standard Deviation**	1735	**p value**	
8	**Sample Size**	100	**Decision**	
9	**Alpha Level**	0.05		
10				
11		**Confidence Interval**		
12		Error	340.05	
13		Lower	Upper	
14		**12159.95**	**12840.05**	

4. Hypothesis Testing. Hypothesis tests determine - if a claim is made - based on numerical information gathered: is it believable? Our study in this section will only include hypothesis tests about a claimed mean.

a. First let's summarize the steps and then discuss each.

- State a claim (hypothesis) about the population
- Select level of risk (α) and determine critical value (z)
- State the decision rule
- Take a representative sample to verify the claim
- Compute sample mean and standard deviation
- Compute z score for sample mean
- Compute p value for sample mean
- Compare z score with critical z (or p value with α)
- Reject or Fail to Reject the claim

b. Step 1 is to make a claim. There are two parts to this process.

(1) First - state the Null Hypothesis (H_o). The null hypothesis is a claim about a population characteristic (the mean) and is assumed to be true. It is also the claim that the test is trying to disprove. In statistics, it is more powerful to reject the null than it is not to. The null hypothesis is usually in the form of and equality, though it can be inclusive of an inequality, i.e. = (equal to), $\leq$ (less than or equal to), or $\geq$ (greater than or equal to).

(2) Next - create an Alternative Hypothesis (H_A or H_1). The alternative is a contradictory claim to the null hypothesis. It is the claim the test is trying to prove. There are three possibilities for the alternative, non-directional, less than, or greater than the claim.

(3) Let's take a look at three examples to illustrate an Alternative Hypothesis.

Example 1. A manufacturer that makes rounds for a 9 millimeter weapon would claim that they are producing rounds that are 9mm in diameter (within some tolerance).

Therefore: H_0: $\mu = 9$ millimeters

H_A: $\mu \neq 9$ millimeters

The alternative in this case is non-directional or not equal to since you don't want rounds larger or smaller than 9mm.

Example 2. A manufacturer claims the cruising range of a new wheeled vehicle will be at least 600 km.

Therefore: H_0: $\mu \geq 600$ kilometers

H_A: $\mu < 600$ kilometers

Example 3. A maintenance shop claims that an item will be at the direct support level no more than 3 days.

Therefore: H_0: $\mu \leq 3$ days

H_A: $\mu > 3$ days

c. Step 2 is to select the level of risk (α).

		True State of Nature	
		H_o is true	H_o is False
Decision	Fail to Reject H_o	Correct decision	Type II error (β)
	Reject H_o	Type I error (α)	Correct decision

First, let's explore what alpha (α) is. When making a decision, there is the true state of nature, or what the truth is. Based on this information, a decision is made about what you believe the truth is. We can make a chart to demonstrate this. There are two correct decisions that can be made, and two errors. You would like to make a correct decision; however, there is always the risk of making an error. Saying the null is not true, when it is true, is called a Type I error. Alpha is the risk you are willing to take to make such an error. Saying the null is true when it is not is called a Type II error. This error, beta

(**β**), is not set by you. Typical settings of alpha are, 0.01, 0.05. and 0.10. The next part of this step is to determine the critical value (**z**) based on the selected alpha level and the alternative hypothesis. A not equal alternative is also called a two tail test because half the error is in each tail, similar to a confidence interval. A less than alternative is called a left tailed test with all the error on the left side. A greater than alternative is called a right tailed test with all the error on the right side. The tail which encompasses the error is called the reject region. The z critical, is the z score, or scores associated with the points, that start the reject region. For example - the z critical for a two tailed test for $\alpha = 0.05$ is the same as for a confidence interval - which is 1.96; however you also have to include the z for the left side which would be – 1.96. A right tail test with $\alpha = 0.05$ would have a different z critical. In this case, all of the error is in the right side and you would look up 0.95 in your table to get a z critical of 1.645. Since a normal distribution is symmetrical, the z critical for a left tailed test with $\alpha = 0.05$ would be – 1.645.

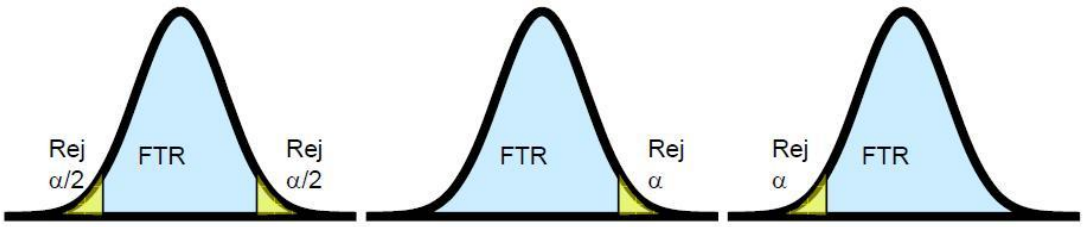
d. Step 3 is to state the decision rule.

In general terms the decision rules follow:

- For a two tailed test: If test statistic > + Z critical or < - Z critical then reject the null hypothesis.
- For a right tailed test: If test statistic > Z critical then reject the null hypothesis.
- For a left tailed test: If test statistic < Z critical then reject the null hypothesis.
- Using p value: If p value < α then reject the null hypothesis.

And conclude: Empirical data strongly suggests the alternative hypothesis is true. Otherwise Fail to reject the null hypothesis and conclude: Insufficient evidence to prove the alternative hypothesis is true.

Below are graphs for a two tailed, right and left tailed test.



e. Step 4 is to gather your sample data.

f. Step 5 is to compute the sample mean and standard deviation as discussed in descriptive statistics.

g. Step 6 is to compute the test statistic **Z** score. The following formula is used for this computation: $T.S. = \dfrac{\bar{x} - \mu_o}{s/\sqrt{n}}$.

h. Step 7 is to compute the **p** value. This step is not actually required, but the p value provides additional information to the analyst. The p value is the lowest level of significance at which the observed value of the test statistic is significant. In other words, how much or little risk can I take to reject the null? After computing the test statistic z score above, look up the z score in the normal distribution table. To compute the p-value:

- For a two tailed test: take $\frac{1-P(z)}{2}$
- For a left tailed test: take P (z)
- For a right tailed test: take 1-P (z)

i. Step 8 is to compare the Test Statistic (T. S.). to the z critical, or the p value to α, and apply the decision rule.

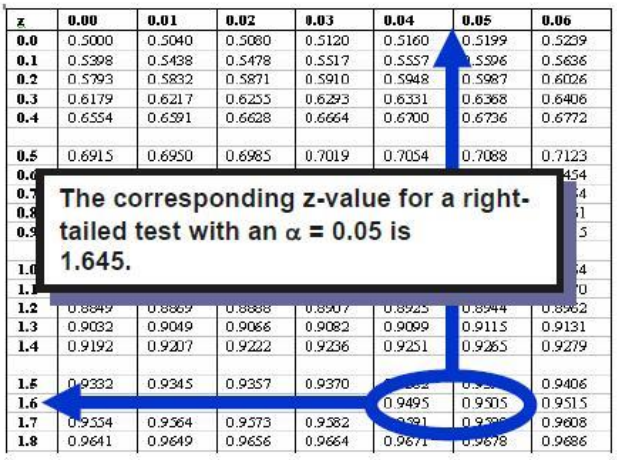
j. Step 9 is to state your conclusion.

k. Let's go back to the example of the vehicle warranties and test the hypothesis that the fleet is traveling on average 12,000 miles. Our Hypothesis would be:

H_o: μ≤12,000 miles

H_A: μ > 12,000 miles

Let's set α = 0.05. Let's look at how we searched for the z critical using a table. Since this is a right-tailed test, we will search for z with a probability of 0.95. Since 0.95 falls exactly between 0.9495 and 0.9505, we went ahead and interpolated to get a z critical of 1.645.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.5	0.9332	0.9345	0.9357	0.9370			0.9406
1.6					0.9495	0.9505	0.9515
1.7	0.9554	0.9564	0.9573	0.9582			0.9608
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686

Our decision rule will then be Reject H_o if T.S. > 1.645.

From our random sample, we obtained the following data:

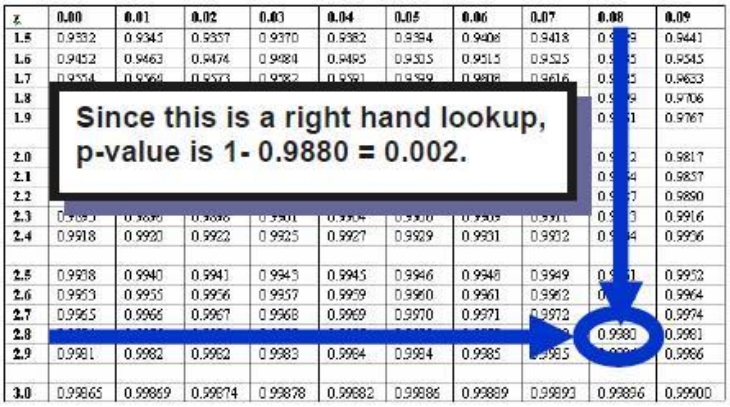
$\bar{x}$ = 12,500 miles

s = 1,735 miles

n = 100 vehicles

We can now calculate our test statistic: $T.S. = \frac{\bar{x}-\mu_o}{s/\sqrt{n}} = \frac{12500-12000}{1735/\sqrt{100}} = 2.88$

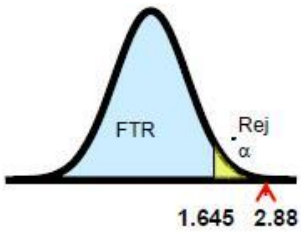
We can also look up our p value:



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9408	0.9418		0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525		0.9545
1.7										0.9633
1.8										0.9706
1.9										0.9767
2.0										0.9817
2.1										0.9857
2.2										0.9890
2.3										0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932		0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949		0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962		0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972		0.9974
2.8									0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985			0.9986
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900

Since this is a right hand lookup, p-value is 1- 0.9880 = 0.002.

Comparing our test statistic we find 2.88 > 1.645. Comparing p value we find 0.002 < 0.05. In both cases we reject H_o and conclude there is sufficient evidence to suggest that our fleet is driving more than 12,000 miles.



What the p value also indicates is that we can set alpha as low as 0.002 and still reject the null. Such low risk gives us a lot of confidence that we are driving more than 12,000 miles.

Normal Distribution

Type Tail Test 1=Two, 2=Left, 3=Right	3		
Hypothesized Mean	12000	Zcritical	1.644854
Sample mean	12500	T.S.	2.881844
Standard Deviation	1735	p value	0.00198
Sample Size	100	Decision	REJECT
Alpha Level	0.05		
	Confidence Interval		
	Error	340.05	
	Lower	Upper	
	12159.95	12840.05	

Let's go back to the template in Excel 2007® and work this problem.

We have now told the template we are working a right tailed test by inserting a 3 in the first box and added the hypothesized mean of 12,000 miles. We are now given the same information we calculated manually to include the decision to make.

5. We have covered a small part of inferential statistics. Other studies in hypothesis testing include:

- Tests about a population variance
- Tests concerning a population proportion
- Tests with paired data
- Tests about differences between population means, variance, and proportions
- Goodness-of-Fit Test
- Test for Independence

You are encouraged, if you do any analysis involving sampling, to continue studying these subjects.

SECTION SIX
REGRESSION
(Return to Table of Contents)

1. Introduction.

   a. Regression analysis is an area of applied statistics which attempts to quantify a relationship among variables and then to describe the accuracy of this established relationship by various indicators. This definition can be divided into two parts. First, quantifying the relationship among the variables involves some mathematical expression. Second, describing the accuracy of the relationship requires the computation of various statistics which will indicate how well the mathematical expression describes the relationship among the variables. Only simple linear regression will be examined, which means that the mathematical expression describing the relationship among the variables will be a linear expression with only two variables and can be graphically represented by a straight line.

   b. The main problem in analyzing bivariate (two variables) and multivariate (more than two variables) data is to discover and measure the association or covariation between (or among) the variables; that is, to determine how the variables vary together. When the relationship between (or among) variables is sharp and precise, ordinary mathematical methods suffice. Algebraic and trigonometric relationships have been successfully studied for centuries. When the relationship is blurred or imprecise, ordinary mathematical methods are not very helpful, but statistical methods are. The special contribution of statistics in this context is that of handling vague, blurred, or imprecise relationships between variables. We can measure whether the vagueness is so great that there is no useful relationship at all. If there is only a moderate amount of vagueness, we can calculate what the best prediction would be and also qualify the prediction to take into account the imprecision of the relationship.

   c. There are two related but distinct aspects of the study of an association between variables. The first, regression analysis, attempts to establish the "nature of the relationship" between variables - that is, to study the functional relationship between the variables and thereby provide a mechanism of prediction, or forecasting. The second, correlation analysis, has the objective of determining the "degree of the relationship" between variables.

2. Correlation Among Variables.

   a. If there is a relationship among any group of variables, there are four possible reasons for this occurrence.

   b. The first and most useless of these reasons is chance. Everyone is familiar with this type of unexpected and unexplainable event. An example of a chance relationship might be a person totally uneducated in the game of football winning a football pool by selecting all the winning teams correctly. This type of relationship is totally useless since it is unquantifiable.

c. A second reason for relationships among variables might be a relationship to a third set of circumstances. For instance, while the sun is shining in the United States, it is night in Australia. Neither event caused the other. The relationship between these two events is better explained by relating each event to another variable, the rotation of the earth with respect to the sun. Although many relationships of this form are quantifiable, a more direct relationship is desired.

d. The third reason for correlation is a functional relationship. These are the relationships that can be represented by equations. An example would be W = m x g where W = weight, m = mass, and g = acceleration due to the force of gravity. Many times this precise relationship does not exist.

e. The last reason considered is the causal type of relationship. These relationships can also be represented by equations, but in this case a cause and effect situation is inferred among the variables. It should be noted that regression analysis does not prove cause and effect. Regression analysis is a tool which allows the analyst to imply that the relationship among variables is consistent. Therefore, two different types of variables will arise.

1) There will be known variables that will be called independent variables and will be designated by the symbol X.

2) There will be unknown variables which will be called dependent variables and will be designated by the symbol Y.

3. Example of Simple Linear Regression Analysis.

In developing cost estimating relationships, simple linear regression analysis will be used most of the time. Although this may seem like an oversimplification of the problem, there are several good reasons for taking this approach. We have discovered that costs often do and should (logically) vary linearly with most physical and performance characteristics, at least over somewhat narrow ranges. If not exactly linear, linear approximations are often adequate. In addition, many curvilinear and exponential functions can be transformed to a linear form, thereby lending themselves to linear analysis. And finally, sample sizes are often so small that use of any other form does not appear justified.

4. Population Simple Linear Regression Model.

a. In a population simple linear regression model, a dependent, or explained, variable Y is related to an independent, or explanatory, variable X by the following expression:

$$\mathbf{Y_i = \beta_0 + \beta_1 X_i + E_i}$$

where i represents the i^{th} coordinated pair of data, β_o (the y intercept of the regression line) and β_1 (the slope of the regression line) are the unknown regression parameters called the population regression coefficients, and E is the random error or residual disturbance term. The dependent variable is Y and the in- dependent variable is X.

b. Designating the variables as dependent or independent refers to the mathematical or functional meaning of dependence; it implies neither statistical dependence nor cause-and- effect. We mean, in the algebraic sense, that we are regarding Y as a function of X and perhaps some other things besides X.

c. It should be noted that the simple linear dependence model, $Yi = \beta_0 + \beta_1 X_i + E_i$, has two distinct parts: the systematic part, $\beta_0 + \beta_1 X_i$, and the stochastic (random) part, E_i. This dissection shows that the model is probabilistic rather than deterministic. The stochastic nature of the regression model implies that the value of Y can never be predicted exactly as in a deterministic case. The uncertainty concerning Y_i is attributed to the presence of E_i. Since E_i is a random variable, it imparts a randomness to Y_i.

5. Assumptions for Simple Linear Regression Analysis.

a. In order to assure the validity of using simple linear regression analysis, several assumptions must be made.

1) First, the independent variables are assumed to be measured without error. This is, in effect, saying that any deviation will be restricted to the dependent variable.

2) The random error terms (E_i) are assumed to be normally distributed about the regression line. In other words the data points are spread evenly both above and below the regression line and there are fewer data points the further you get from the regression line. This assumption allows us to use certain statistics and tests of significance to evaluate the validity of the regression line.

3) The random error term is assumed to come from an identically and independently distributed variable with mean zero and constant variance. This means that the random error term cannot be predicted from a knowledge of the independent variable, X.

b. Using the above assumptions, estimators for the unknown regression parameters, β_0 and β_1, can be derived and inferences using these estimators can be made. It should be stressed that, in practice, one or more of the assumptions are often violated. Frequently, the independent variables are not measured without error, and the sample size is so small that assumptions about normality are invalid. Unless these assumptions are grossly violated, the bias in estimating β_0 and β_1 should be minimal.

6. Sample Simple Linear Regression Model.

   a. As you may have noticed, the population regression model represents the equation of a straight line. Since sample data will virtually always be used (it is rare when all points in the population are known or identified), the model for the sample data is:

   $$\hat{\mathbf{y}} = \mathbf{b_0} + \mathbf{b_1 x}$$

   where $\hat{\mathbf{y}}$ is the predicted value of Y,
   $\mathbf{b_0}$ is the predicted value of β_0,
   $\mathbf{b_1}$ is the predicted value of β_1, and
   x is the independent variable.

   b. All three of these predictors are unbiased estimators of their predicted values.

7. Evaluating the Coefficients.

   a. As can be seen, the model is an extremely simple relationship. Still, evaluating b_O and b_1 does require a rigorous mathematical approach. The method used to evaluate b_O and b_1 is known as the method of least squares. This method fits a line to the data points that minimizes or makes as small as possible the sum of the squares of the vertical distances of the data points from the line. In order to use this approach, the problem must be better defined and this can best be done graphically.

   b. Referring to the figure on the next page, suppose x_i represents the i^{th} value of the independent variable.

      1) The value of the dependent variable $\mathbf{y_i}$ is the actual observed value associated with $\mathbf{x_i}$.
      2) $\hat{\mathbf{y}}_\mathbf{i}$ is the predicted value associated with x_i (i.e. the result of solving the regression equation for the given value X_i). $\hat{\mathbf{y}}_\mathbf{i}$ will differ from $\mathbf{y_i}$ whenever $\mathbf{y_i}$ does not lie on the regression line.
      3) $\bar{\mathbf{y}}$ is the mean value of all the $\mathbf{y_i}$ values. Note that the point $\mathbf{(\bar{x}, \bar{y})}$ will always lie on the regression line.

   c. Refer again to the figure on the next page. $\bar{\mathbf{y}}$ serves as a baseline for all the Y values. If you were asked to predict a value for y given a value for x and you didn't have any other information, a reasonable guess would be the mean $\bar{\mathbf{y}}$. In all likelihood the actual value of y will not correspond to the mean, so the difference between the mean and the actual value of y is the very worst you should do. This deviation or the difference between **y** and $\bar{\mathbf{y}}$ is known as the total deviation. This deviation can be separated into two parts. The regression equation provides us with a predicted value of **y** (i.e. $\hat{\mathbf{y}}$), which is closer to the actual value of **y**. The

difference between this predicted value and the mean (i.e. $\hat{y} - \bar{y}$) is called the explained deviation because it is accounted for by the regression line. The remaining deviation, the difference between the actual value and predicted values (i.e. $y - \hat{y}$) is called the unexplained deviation. In other words:

Total deviation = $y - \bar{y}$
xplained deviation = $\hat{y} - \bar{y}$
nexplained deviation = $y - \hat{y}$
and
$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$

Including all of the data points results in the following equation:

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

The sum of squares is taken over all the data points and is now referred to as variation and not deviation. If you sum together all of the vertical deviations from the regression line, the result would be zero. To eliminate this problem the individual deviations are squared.

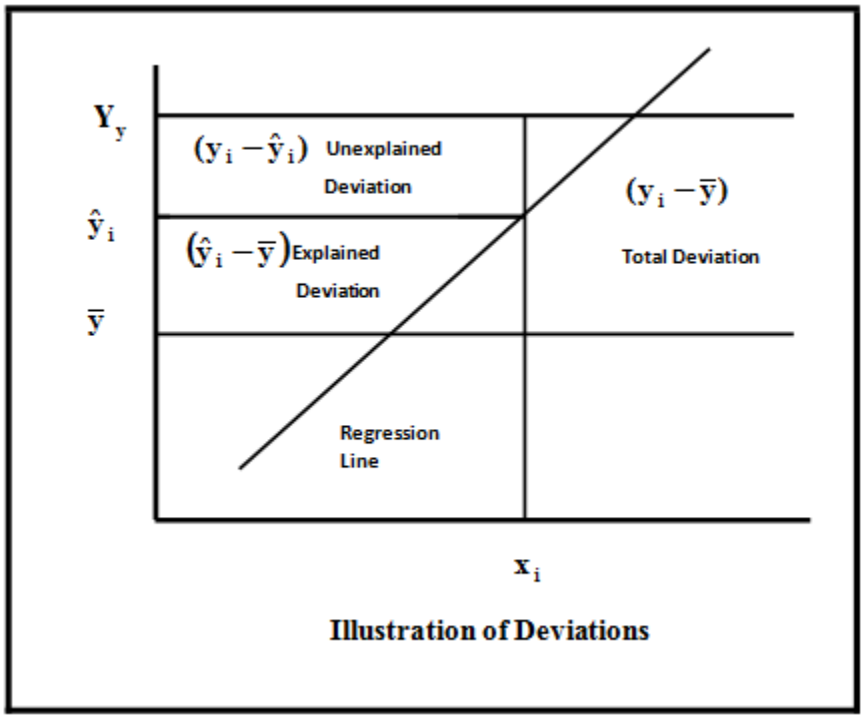


Illustration of Deviations

d. So simple linear regression uses the method of least squares to fit a line to the sample data points that minimizes the sum of the squares of the unexplained deviations. This criterion makes the regression line as close to all the points as is possible. Therefore, it is desired to minimize

$$\sum (y_i - \hat{y}_i)^2$$

It is now necessary to determine the least squares estimators for b_0 and b_1. The equation to find b_1 is shown below.

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Performing some algebra on this equation results in equation shown below which is actually easier to use if you must calculate b_1 manually.

$$b_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

The y-intercept $\mathbf{b_0}$ can be found using the equation $b_0 = \frac{\sum y - b_1 \sum x}{n}$.
Since the ordered pair $(\bar{x}, \bar{y})$ always lies on the regression line $\mathbf{b_0}$ can be found by substituting the values for $\mathbf{b_1}$, $\bar{x}$, and $\bar{y}$ into the equation $\hat{y} = b_0 + b_1 x$ and solving for $\mathbf{b_0}$.

e. Now that $\mathbf{b_0}$ and $\mathbf{b_1}$ have been evaluated, the regression line is fully defined and a predicting equation has been developed. Therefore, for any value of the independent variable a predicted "average" value of the dependent variable can be determined by using the regression equation $\hat{y} = b_0 + b_1 x$. Given the assumptions of the simple linear regression analysis, we can assume that for any specified value of the independent variable there can be many different values for the dependent variable. These values for the dependent variable are assumed to be normally distributed about the regression line with the mean value located on the line. Thus, the predicted value for y determined from the regression equation is mean value or "average" value of y associated with the specified value of x.

f. At this point an example that will be used throughout the remainder of this discussion will be introduced. The government is considering buying a newly developed vehicle that weights approximately 2.2 tons. Data is available on ten systems which have been procured in the past giving the weight and first unit cost of each of these vehicles. Therefore a regression line can be computed and an estimate of the first unit cost of the new vehicle can be made based upon its weight.

g. The data for the ten systems are:

x_i (tons)	y_i ($K)	$x_i y_i$	x_i^2	y_i^2
1.0	6	6.0	1.00	36
1.5	11	16.5	2.25	121
2.0	14	28.0	4.00	196
1.8	16	28.8	3.24	256
1.9	20	38.0	3.61	400
2.5	18	45.0	6.25	324
3.0	22	66.0	9.00	484
3.4	30	102.0	11.56	900
3.7	26	96.2	13.69	676
3.9	31	120.9	15.21	961
$\sum x = 24.7$	$\sum y = 194$	$\sum xy = 547.4$	$\sum x^2 = 69.81$	$\sum y^2 = 4354$

Therefore, substituting the above values into the equations for $\mathbf{b_1}$ and $\mathbf{b_0}$ we can determine the parameters of the regression equation.

$$\mathbf{b_1 = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} = \frac{10(547.4) - (24.7)(194)}{10(69.81) - (24.7)^2} = 7.751392}$$

$$\mathbf{b_0 = \frac{\sum y - b_1 \sum x}{n} = \frac{194 - 7.751(24.7)}{10} = 0.25406}$$

The y-intercept could have been found by substituting the values for $\mathbf{b_1}$, $\bar{x}$, and $\bar{y}$ into the equation $\hat{y} = b_0 + b_1 x$ and solving for $\mathbf{b_0}$ as shown below.

$$\bar{x} = \frac{24.7}{10} = 2.47 \qquad \bar{y} = \frac{194}{10} = 19.4$$

$$\hat{y} = b_0 + b_1 x$$

$$b_0 = \hat{y} - b_1 x = 19.4 - (7.751392)(2.47) = 0.25406$$

The regression equation then becomes $\hat{y} = 0.254 + 7.751x$.

The predicted "average" first unit cost for the new system is

$$\hat{y} = 0.254 + 7.751x = 0.254 + 7.751(2.2) = \$17.307K = \$17{,}307.$$

h. Regression analysis coefficients are sensitive to rounding errors. The analyst cannot expect these errors to somehow cancel each other. As a rule of thumb, always carry intermediate calculations at least two (or more) decimal places further than the number of places desired in the final answer. The reason that regression analysis is especially subject to rounding error is that, often, two large numbers must be calculated and their difference calculated. If the difference is small, and often it is, the difference between rounded numbers may be meaningless.

i. The regression coefficients $\mathbf{b_0}$ and $\mathbf{b_1}$ may or may not have a practical meaning depending on the problem. In the previous problem weight was used to predict vehicle cost. The value of $b_0 = 0.254$ implies that a vehicle with a weight of zero tons would cost \$254, which is illogical. One should realize that a weight of 0 tons is outside of the range of the data. Extrapolating outside the range of the data is dangerous because we can make no inference as to the linearity of the data beyond the sample range. In this case we may say that as vehicle weight is reduced beyond a certain point, cost may increase.

j. The regression slope $\mathbf{b_1}$ is of both theoretical and practical importance. Theoretically, together with $\mathbf{b_0}$ the position of the regression line can be determined by the slope. Also the value of $\mathbf{b_1}$ will be used to test the significance of the total regression. Practically, the slope measures the average change in the predicted value of y, for a one unit change in x. From the previous example the slope of the regression line was 7.751 (i.e. $b_1 = 7.751$). This implies that the cost of a system will increase by \$7751 each time the vehicle's weight increases by one ton.

k. The discussion thus far has centered about quantifying the relationship between independent and dependent variables. This, as previously mentioned, constitutes only one half of the regression analysis problem. Statistics and tests of significance which measure the validity of the regression line need now to be computed.

8. Correlation Analysis and Regression Statistics.

a. Coefficient of Determination: The coefficient of determination is a measure of how good the least squares line is an instrument of regression or to put it another way, it is a measure of how good the regression line fits the data. This coefficient of determination, represented symbolically by $\mathbf{r^2}$, is the ratio of explained variation over total variation (refer back to the figure Illustration of Deviation). In other words, it is a measure of the proportion of variation in y that is explained by the regression line, using x as the explanatory variable.

$$\mathbf{r^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{Explained\ variation}{Total\ variation}}$$

The coefficient of determination, being a proportion, ranges in value from 0 to 1. You'd like to achieve an r^2 value as large as possible. Acceptable levels depend upon the problem situation and may range anywhere from 0.5 or above. In cost estimating, for example, an acceptable r^2 value is usually greater than or equal to 0.9.

Within simple linear regression the coefficient of determination can be used to compare various regression lines to determine the best prediction model. Select the regression model with the largest r^2 value.

There is a computational formula that makes the manual calculation easier.

$$\mathbf{r^2 = \frac{\left(\sum x_i y_i - n\bar{x}\bar{y}\right)^2}{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}$$

The coefficient of determination for the above regression example can be determined as shown below.

$$\mathbf{r^2 = \frac{\left(\sum x_i y_i - n\bar{x}\bar{y}\right)^2}{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)} = \frac{\left[547.4 - 10(2.47)(19.4)\right]^2}{\left[69.81 - 10(2.47)^2\right]\left[4354 - 10(19.4)^2\right]} = 0.8957}$$

A coefficient of determination of 0.8957 implies that 89.57% or about 89% of the variation in first unit costs can be explained by the variation in vehicle weight, leaving about 11% to chance or other variables.

b. Pearson's Product Moment Correlation Coefficient: The Pearson correlation coefficient, designated by **r**, is a numerical measurement that assesses the strength of a linear relationship between two variables x and y. It has the following characteristics:

1) r is a unitless measurement between – 1 and 1. In symbols $-1 \leq r \leq 1$. If r = 1, there is a perfect positive linear correlation and if r = – 1, there is a perfect negative correlation. If r = 0, there is no linear correlation. The closer r is to 1 or – 1, the stronger the correlation and the better a straight line describes the relationship between the two variables x and y.

2) Positive values of r imply that as x increases, y tends to increase. Negative values of r imply that as x increases, y tends to decrease. With respect to the regression line, a line with a positive slope has a positive r and a line with a negative slope results in a negative value for r.

3) The value of r is the same regardless of which variable is the explanatory variable and which is the response variable. In other words, the value of r is the same for the pairs (x, y) and the corresponding pairs (y, x).

4) The value of r does not change when either variable is converted to different units of measure.

The correlation coefficient **r** has the following formula:

$$\mathbf{r = \frac{1}{n-1}\sum \frac{(y-\bar{y})}{s_y} \cdot \sum \frac{(x-\bar{x})}{s_x}}$$

An easier computational formula when calculating **r** manually is shown below.

$$\mathbf{r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}}$$

Referring back to the regression example relating first unit cost to vehicle weight, the calculations for the correlation coefficient are shown below.

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}} = \frac{10(547.4) - (24.7)(194)}{\sqrt{10(69.81) - (24.7)^2}\sqrt{10(4354) - (194)^2}}$$

$r = 0.9464$

(There is a strong positive relationship between vehicle weight and first unit costs.)

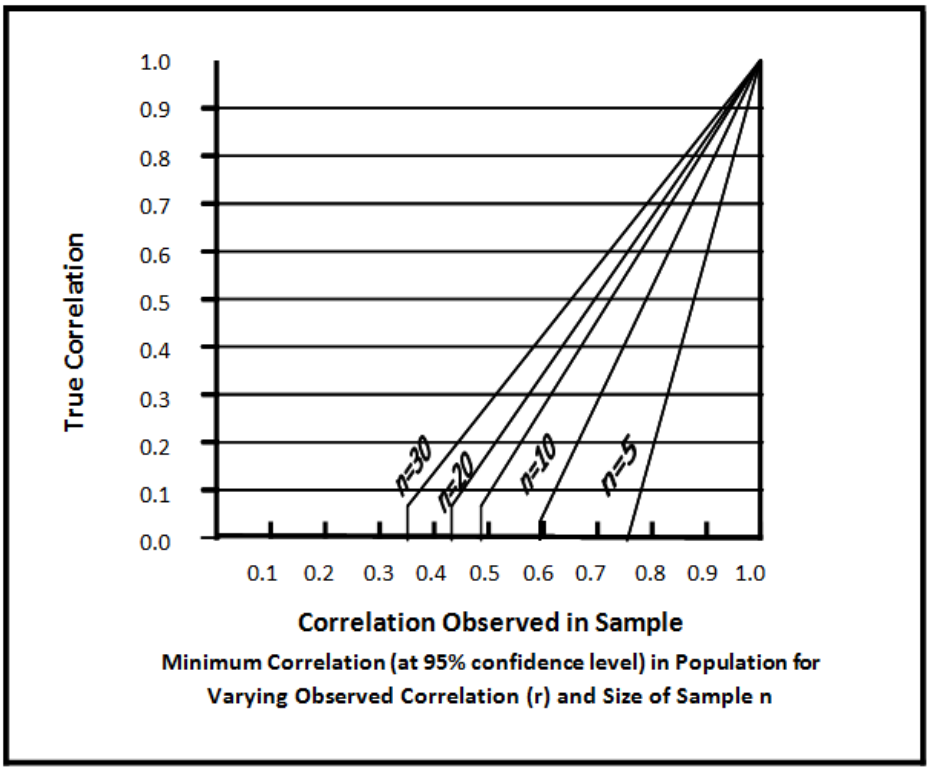
There is a relationship between the Pearson correlation coefficient, **r**, and the coefficient of determination, $\mathbf{r^2}$. The correlation coefficient is the positive square root of the coefficient of determination multiplied by the slope of the regression line divided by the absolute value of the slope. Symbolically,

$$\mathbf{r = \frac{b_1}{|b_1|}\sqrt{r^2}}$$

The multiplication of the positive square root by $\frac{b_1}{|b_1|}$ results in the correct sign on r.

In the regression example, $\mathbf{r = \frac{b_1}{|b_1|}\sqrt{r^2} = \frac{7.751}{|7.751|}\sqrt{0.8957} = 0.9464}$, which is the same value as found above.

The correlation coefficient is often misinterpreted. A high value for r does not prove cause –and-effect. It should also be noted that there is a sampling error associated with the correlation coefficient. This is particularly true when working with small sample sizes. For samples of size 10 and 5, r values less than 0.60 and 0.80 respectively are meaningless (i.e. at the 95% confidence level, it cannot be said that the true value of r is significantly different from zero). The figure below reflects the minimum value of the population correlation coefficient associated with various sample sizes and sample correlation coefficients.



c. Standard Error of the Estimate: The standard error of the estimate, designated by $\mathbf{S_e}$, is a measure of the spread of the data points about the regression line. It is analogous to the sample standard deviation of measurements of a single variable. The definitional formula for finding $\mathbf{S_e}$ is

$$\mathbf{S_e} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

The size of $\mathbf{S_e}$ is dictated by the size of the residuals (i.e. the difference between the actual data value and the predicted value). The nearer the scatter points lie to the regression line, the smaller $\mathbf{S_e}$ will be and the more scattered the points about the line, the larger $\mathbf{S_e}$ will be. In fact, if $S_e = 0$, then all the points would lie on the regression line with there being no difference between predicted and actual values.

The standard error of the estimate will have the same units of measurement as the dependent variable. It should not be used to compare different regression lines. It is used primarily to calculate other statistics and tests of significance. It is desirable that **Se** be as small as possible. Calculating $\mathbf{S_e}$ using the definitional formula shown above can be rather tedious. A mathematically equivalent computational formula is shown below.

$$\mathbf{S_e} = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n - 2}}$$

The standard error of the estimate for the example regression problem is

$$S_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}} = \sqrt{\frac{4354 - 0.254(194) - 7.751(547.4)}{10-2}} = \$2.78K$$

d. Coefficient of Variation: The coefficient of variation is the statistic which allows the use of the standard error of the estimate to compare different regression lines. It is actually a relative standard error of the estimate since it becomes a dimensionless quantity. The symbol for the coefficient of variation is $\mathbf{C_V}$, and the formula is

$$\mathbf{C_v = \frac{S_e}{\bar{y}}}$$

$\mathbf{C_v}$ measures $\mathbf{S_e}$ as a percentage of $\mathbf{\bar{y}}$. When developing a cost estimating relationship using regression analysis, it is desired that $\mathbf{C_V}$ be less than or equal to 0.1.

The coefficient of variation for the example problem is

The coefficient of variation for the example problem is $\mathbf{C_v = \frac{S_e}{\bar{y}} = \frac{2.78}{19.4} = 0.143.}$

9. Inferences About Population Regression Coefficients (Hypothesis Tests).

a. Having obtained the sample regression equation, and having concluded that the regression equation may be a useful one on the basis of the sample standard error of estimate ($\mathbf{S}_e$) and the coefficient of determination ($\mathbf{r^2}$), the analyst might assume the equation to be a valid predictive device. However, the equation may still contain some error because the relationship between x and y is not perfect. Predicting may not be precise because of sampling error, or chance variations in sampling. Because of the assumption previously made concerning normally distributed error terms, certain hypothesis tests for the significance of the intercept and the slope can be performed.

b. From a practical standpoint, the significance of the intercept is of little importance; it is usually outside the range of the data. On the other hand, the significance of the value of the slope, b_1 is great. The hypothesis test might be used to determine if the slope was significantly different from any value. A regression line with slope 0 is horizontal, which implies that the mean of y does not change at all when x changes. A slope of 0 implies that there is no true linear relationship between x and y or that the regression equation has no value for predicting y. If the slope of the regression line is significantly different from 0, then the regression equation can be used to predict y.

c. In order to perform a hypothesis test about the slope of the regression line, another statistic, the standard error of the slope must be computed. The formula for computing the standard error of the slope is

$$\mathbf{SE_b = \frac{S_e}{\sqrt{\sum x^2 - n\bar{x}^2}} = \frac{S_e}{\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}}}$$

d. The procedure for conducting the hypothesis test about the slope of the regression line is as follows:

1) Establish the null and alternative hypotheses.

$$\mathbf{H_0: \beta_1 = 0}$$
$$\mathbf{H_1: \beta_1 \neq 0}$$

2) Determine the desired level of significance, α, for the test.

3) Use the Student t distribution (see Appendix D) to find $\mathbf{t_{\alpha/2,\, n-2}}$.

4) Calculate the test statistic

$$\mathbf{t = \frac{b_1 - \beta_1}{SE_b} = \frac{b_1}{SE_b}}$$ since $\beta_1 = 0$ (see the null hypothesis above).

5) Conclude the test.

a) **Reject $H_0: \beta_1 = 0$ if $|t| \geq t_{\alpha/2,n-2.}$** and conclude that the slope of the regression line is significantly different from 0. This implies that the regression equation can be used to predict y.

b) **Fail to reject $H_0: \beta_1 = 0$ if $|t| \leq t_{\alpha/2,n-2.}$.** Since the null hypothesis cannot be rejected, there is insufficient evidence to say that the slope is significantly different from 0. The regression equation should not be used to predict y.

e. Following the algorithm given above, the example problem will yield the following results.

1) The null and alternative hypotheses are

$$\mathbf{H_0: \beta_1 = 0}$$
$$\mathbf{H_1: \beta_1 \neq 0}$$

2) The level of significance is $\alpha = 0.05$ (chosen arbitrarily).

Find the critical value from the Student t distribution in Appendix D.

$$\mathbf{t_{\alpha/2,\,n-2} = t_{0.025,\,8} = 2.306}$$

3) Calculate the standard error of the slope.

$$SE_b = \frac{S_e}{\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}} = \frac{2.78}{\sqrt{69.81 - \frac{1}{10}(24.7)^2}} = 0.937$$

4) Calculate the test statistic.

$$\mathbf{t = \frac{b_1}{SE_b} = \frac{7.751}{0.937} = 8.271}$$

5) Conclude the test and make a decision. Since the test statistic is greater than the critical value from the Student t distribution (i.e. 8.271 > 2.306), reject the null hypothesis. The slope is significantly different from 0 and the regression equation can be used to predict y.

10. Confidence Interval for β.

a. The confidence interval for β is given by

$$\mathbf{b_1 - E < \beta < b_1 + E}$$

$$\mathbf{E = t_{\alpha/2,n-2} SE_b}$$

b. A 95% confidence interval for the example problem is given below.

$$\mathbf{t_{\alpha/2,n-2} = t_{0.025,8} = 2.306}$$
$$\mathbf{SE_b = 0.937}$$
$$\mathbf{E = t_{0.025,8} SE_b = 2.306(0.937) = 2.161}$$

$$\mathbf{b_1 - E < \beta < b_1 + E}$$
$$\mathbf{7.751 - 2.161 < \beta < 7.751 + 2.161}$$
$$\mathbf{5.590 < \beta < 9.912}$$

11. Confidence and Prediction Intervals for a Regression Response.

a. One of the most common reasons to fit a line to data is to predict the response to a particular value of the explanatory variable. The method is simple enough: just substitute the value of the explanatory variable x into the regression equation and solve for the value of the response variable y.

A confidence interval is required to describe how accurate this prediction is, but before this interval can be determined a key question must be answered. It is desirable to predict the "mean" response to the given value of x or is it only desirable to predict "a single" response to the given value of x. For example, the regression equation developed above can be used to predict the first unit cost of a newly developed vehicle weighing 2.2 tons. The key question then is this to be a prediction for the mean first unit cost of all vehicles weighing 2.2 tons or the prediction of the first unit cost for a single vehicle weighing 2.2 tons. The actual single point prediction will be the same, but the margins of error will be different.

b. To estimate the "mean" response, use a confidence interval. It is an ordinary confidence interval for the parameter

$$\mu_y = \beta_0 + \beta_1 x_0$$

The regression model says that μ_y is the mean of the responses **y** when **x** has the value x_0. It is a fixed number whose value is unknown.

c. To estimate an individual response **y**, use a prediction interval. A prediction interval estimates a single random response **y** which is not a fixed number. Several observations with $x = x_0$ will result in different responses.

d. The interpretations of the confidence interval and prediction interval are the same. A 95% prediction interval, like a 95% confidence interval is right 95% of the time in repeated use. The main difference between the two is that it is harder to predict one single response than to predict a mean response. Both intervals have the same form

$$\hat{y} - t_{\alpha/2,n-2} SE < y < \hat{y} + t_{\alpha/2,n-2} SE$$

The prediction interval is wider than the confidence interval.

e. A $1 - \alpha$ **confidence interval for the mean response μ_y** when $x = x_0$ is

$$\hat{y} - t_{\alpha/2,n-2} SE_{\hat{\mu}} < \mu_y < \hat{y} + t_{\alpha/2,n-2} SE_{\hat{\mu}}$$

$$SE_{\hat{\mu}} = S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x^2 - n\bar{x}^2}} = S_e \sqrt{\frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - \left(\sum x\right)^2}}$$

f. A $1 - \alpha$ **prediction interval for a single observation on y** when $x = x_0$ is

$$\hat{y} - t_{\alpha/2,n-2} SE_{\hat{y}} < y < \hat{y} + t_{\alpha/2,n-2} SE_{\hat{y}}$$

$$SE_{\hat{y}} = S_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x^2 - n\bar{x}^2}} = S_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n\sum x^2 - \left(\sum x\right)^2}}$$

Returning to the example problem, calculate a 95% confidence interval for the mean first unit cost when x = 2.2 tons.

$$\hat{y} = 0.254 + 7.751x = 0.254 + 7.751(2.2) = 17.307.$$

$$t_{\alpha/2,n-2} = t_{0.025,8} = 2.306 \text{ (from Student t distribution)}$$

$$S_e = 2.78 \text{ (see paragraph 8c above)}$$

$$SE_{\hat{\mu}} = S_e\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x^2 - n\bar{x}^2}} = 2.78\sqrt{\frac{1}{10} + \frac{(2.2 - 2.47)^2}{69.81 - 10(2.47)^2}} = 0.9148$$

$$\hat{y} - t_{\alpha/2,n-2}SE_{\hat{\mu}} < \mu_y < \hat{y} + t_{\alpha/2,n-2}SE_{\hat{\mu}}$$

$$17.307 - 2.306(0.9148) < \mu_y < 17.307 + 2.306(0.9148)$$

$$15.197 < \mu_y < 19.191$$

g. Calculate a 95% prediction interval for the first unit cost of the next newly developed vehicle to weigh 2.2 tons.

$$\hat{y} = 17.307 \qquad t_{\alpha/2,n-2} = t_{0.025,8} = 2.306 \qquad S_e = 2.78$$

$$SE_{\hat{y}} = S_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x^2 - n\bar{x}^2}} = 2.78\sqrt{1 + \frac{1}{10} + \frac{(2.2 - 2.47)^2}{69.81 - 10(2.47)^2}} = 2.9266$$

$$\hat{y} - t_{\alpha/2,n-2}SE_{\hat{y}} < y < \hat{y} + t_{\alpha/2,n-2}SE_{\hat{y}}$$

$$17.307 - 2.306(2.9266) < \mu_y < 17.307 + 2.306(2.9266)$$

$$10.558 < y < 23.336$$

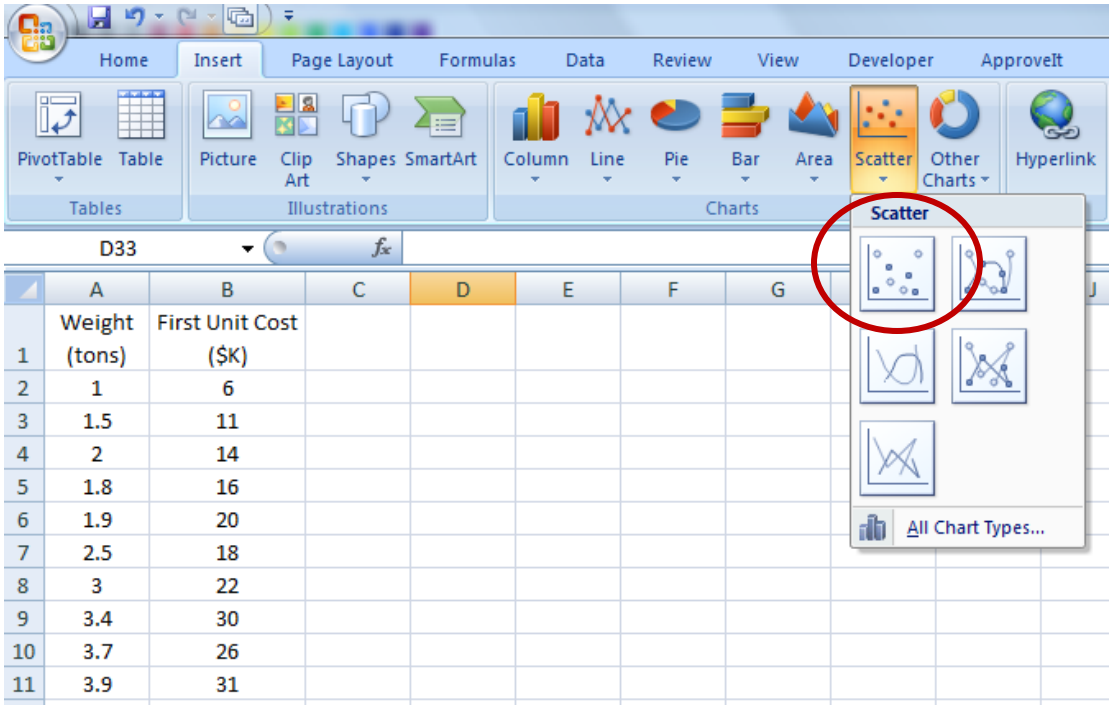
h. From the above we can conclude that we are 95% confident that on the average a vehicle weighing 2.2 tons will have a first unit cost of somewhere between \$15.197K and \$19.191K. However, if we are trying to predict the first unit cost of a single vehicle weighing 2.2 tons, we are 95% confident that this first unit cost will be somewhere between \$10.558K and \$23.336K. Note that the prediction interval for the prediction of a single response is wider than the confidence interval for the mean response.

12. Regression in Excel.

a. Using the previous example, let's now see how we can use Excel for simple linear regression. The first thing we should always do is graph our data. This will give us an indicator of whether regression is appropriate and an idea of what to look for in the calculations. First type the data into Excel.

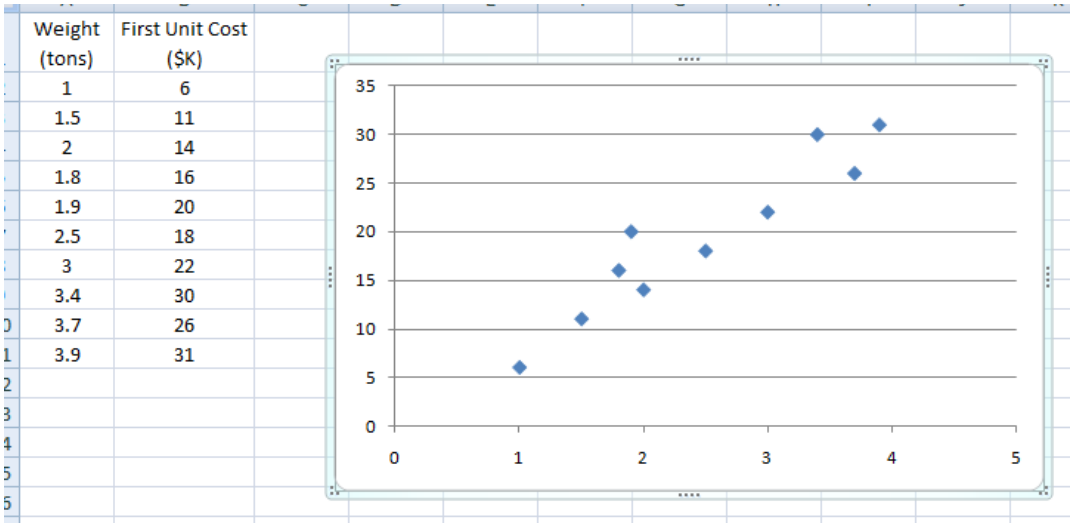
	A	B
1	Weight (tons)	First Unit Cost ($K)
2	1	6
3	1.5	11
4	2	14
5	1.8	16
6	1.9	20
7	2.5	18
8	3	22
9	3.4	30
10	3.7	26
11	3.9	31

Then click on the Insert Ribbon tab to reveal the Insert Ribbon. Highlight the data and left click on Scatter Plot to reveal the Scatter Plot menu as shown on the figure at the top of the next page. Then select the upper left-hand scatter plot.



Selecting the upper left-hand scatter plot (**circled above**) will result in the scatter

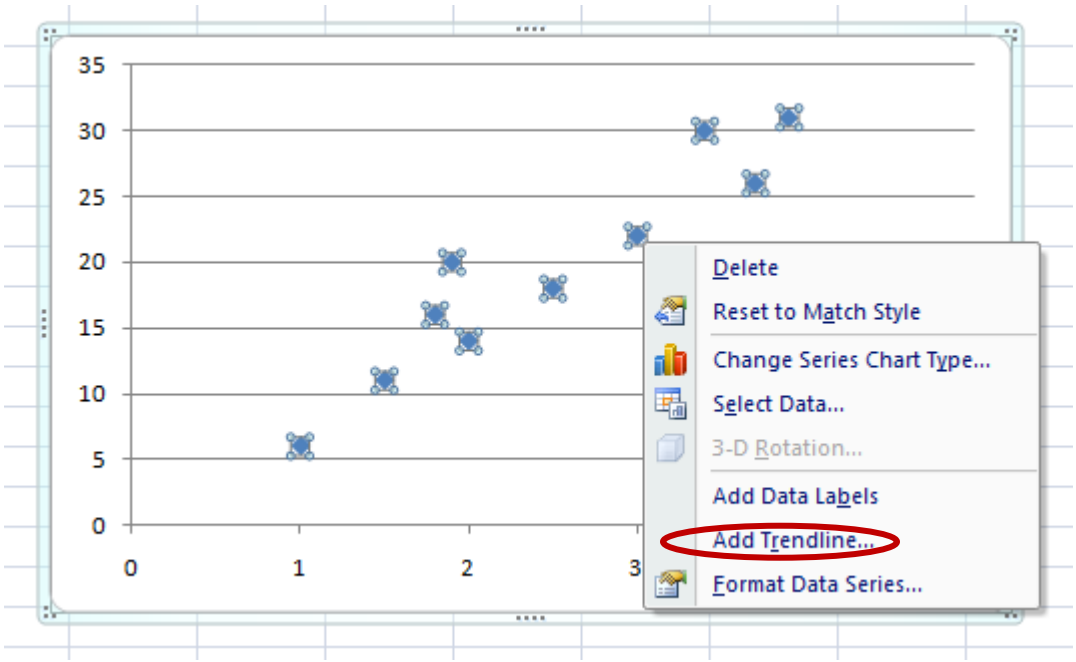
plot shown below.



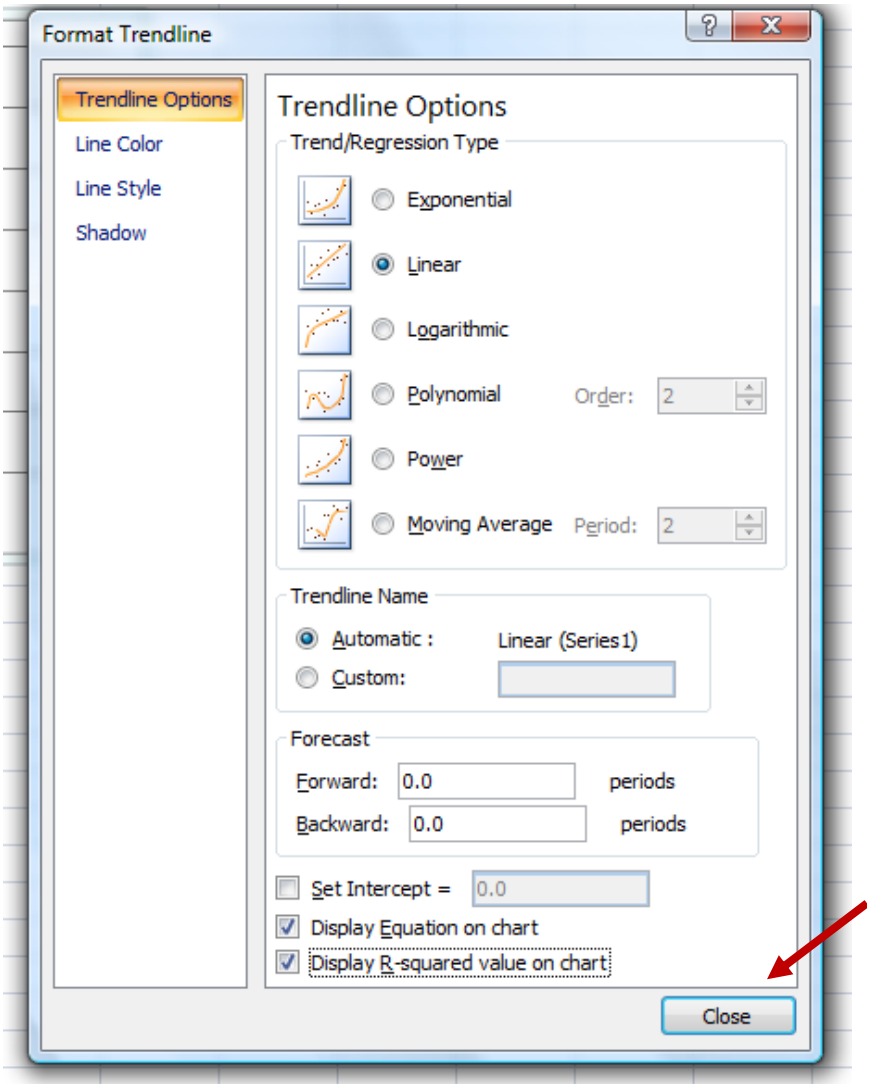
Weight (tons)	First Unit Cost ($K)
1	6
1.5	11
2	14
1.8	16
1.9	20
2.5	18
3	22
3.4	30
3.7	26
3.9	31

This scatter plot indicates that there is a strong positive linear relationship between weight and first unit cost.

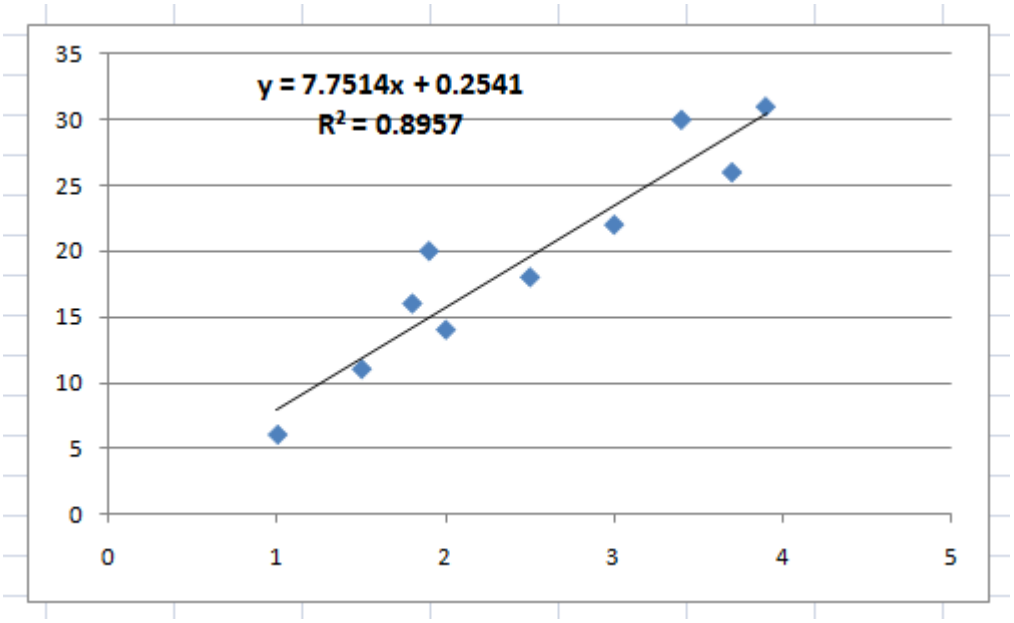
b. We will examine two ways to find the regression equation. The first is to add a trend line to the scatter plot. Right click on any of the data points to reveal the menu shown below.



Right click on Add Trendline (circled in the above figure) to reveal another menu displaying trendline options. Insure that linear is marked and then check the boxes for both "Display Equation on chart" and "Display R-squared value on chart" as shown in the figure on the next page.

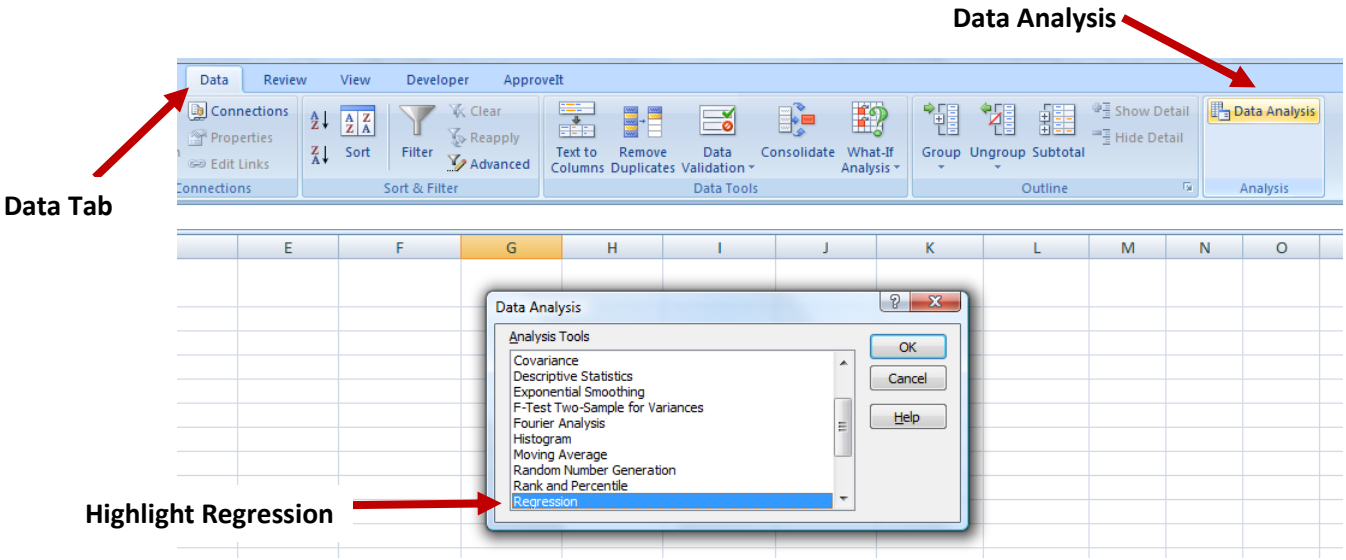


Click on the close button to add the trendline to the scatter plot and display the equation and r^2 value on the graph as displayed on the scatter plot shown on the next page.

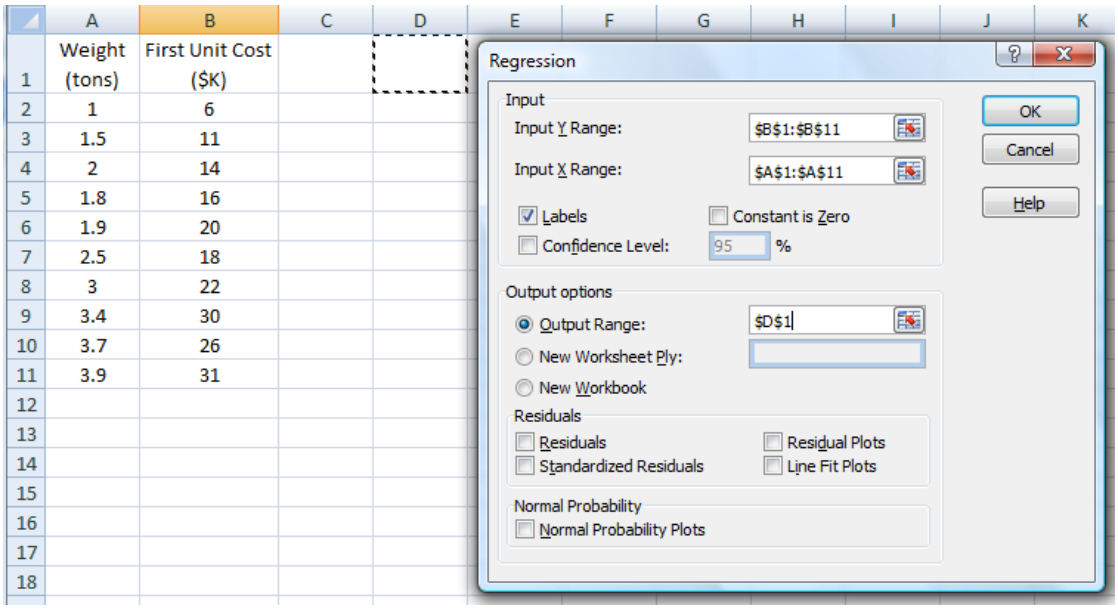


The equation and r^2 value agree with what we previously calculated. This regression model accounts for about 89.6% of the variation in first unit cost.

c. Another way of gathering the calculations is by creating an Analysis of Variance (ANOVA) table. To do this, we will use the data analysis add-in (refer to Appendix E for a discussion of the installation of Add-ins in Excel). Left click on the Data tab to reveal the Data Ribbon. Then left click on Data Analysis (far right-hand side of the ribbon) to bring up the menu shown below. Highlight Regression and click OK.



d. Highlighting regression and clicking OK provides the menu shown below. Input the range for the response value y (in this case B1:B11). Input the range of the explanatory variable x (in this case A1:A11). Since these included the column labels for the variables (i.e. Weight and First Unit Cost) check labels. Click on output range and identify a reference cell (in this case D1). This cell locates the upper left-hand corner for the output. Then click on OK.



e. Clicking on OK results in a display of the output data.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.946395203
R Square	0.89566388
Adjusted R Square	0.882621865
Standard Error	2.774888409
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	528.7999546	528.7999546	68.6752681	3.38516E-05
Residual	8	61.60004545	7.700005681		
Total	9	590.4			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.254062038	2.471373192	0.102801972	0.92065036	-5.444934756	5.953058833	-5.444934756	5.953058833
Weight (tons)	7.751391887	0.935361548	8.287054248	3.38516E-05	5.594444292	9.908339483	5.594444292	9.908339483

Look at the Regression Statistics section of the report (recreated below). It provides the value of r^2 and the standard error of the estimate (Standard Error). Notice that these are the same values we calculated above.

Regression Statistics	
Multiple R	0.946395203
R Square	0.89566388
Adjusted R Square	0.882621865
Standard Error	2.774888409
Observations	10

Now examine the bottom section of the report (recreated below).

y-intercept

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.254062038	2.471373192	0.102801972	0.92065036	-5.444934756	5.953058833
Weight (tons)	7.751391887	0.935361548	8.287054248	3.38516E-05	5.594444292	9.908339433

Slope

The Coefficients column identifies the regression equation parameters b_0 and b_1 which can be used to write the equation $\hat{y} = 0.254 + 7.751x$. Note that this is the same equation we derived earlier.

Look at the "t Stat" value for weight (i.e. t = 8.287). This is the test statistic for the hypothesis test (H_0: $\beta_1 = 0$) and can be compared to the critical value derived from the Student t distribution (2.306 for this hypothesis test) to conclude that the slope of the regression line is significant. Likewise, examine the P-value = 3.385×10^{-5}. Since this is less than the established level of significance, $\alpha = 0.05$, can conclude that the slope is significant.

This section also provides the lower and upper bounds for the 95% confidence interval about the slope of the regression line (see the lower and upper bounds for Weight).

Now examine the ANOVA section of the output report (shown below). Focus on the circled "Significance F" value. This is the P-value and given that it is less than the established $\alpha = 0.05$ level of significance, we can conclude that the regression model is significant and can be used to predict y.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	528.7999546	528.7999546	68.67526831	3.38516E-05
Residual	8	61.60004545	7.700005681		
Total	9	590.4			

13. Conclusion.

a. Regression analysis, while a very useful device, often opens up many opportunities for misinterpretation. The most unwarranted interpretation stems from the confusion between association and causation (regression analysis shows the degree of association). Many people take the independent variable to be the cause and the dependent variable to be the effect; this need not be true.

b. The link between association and causation can be stated: the presence of association does not imply causation; but causation always implies association. Statistical evidence can only show the degree of association between variables. Whether causation exists or not depends purely on reasoning (logic). For example, there is reason to believe that an increase in the weight of a modern tank "causes" an increase for cost (a direct or positive association). There is also reason to support the premise that a decrease in the weight of an aircraft (range, velocity and payload held constant) "causes" an increase in cost (an inverse of negative association). The fact that inches of padding in the commander's seat of a tactical vehicle is associated with high cost does not show cause and effects.

c. Why is it that association does not show causation? One reason is that the association between variables may be pure chance, such as soil erosion in Alaska and the amount of alcohol consumed in South America. Another reason is that association between two variables may be due to the influence of a third common variable. Since 1945, for example, it has been found that there is a close relationship between teacher's salaries and liquor consumption. A plausible explanation is that both variables have been influenced by the continuous increase in national income during the same period. A third possibility is that in the real relationship we may be able to determine which variable is the cause and which the effect. For example, the higher the per capita income is in a state, the more money spent for each student in public education; the more money spent for education, the higher will be the per capita income; which is the cause and which is the effect?

d. Erroneous conclusions may also result from extrapolating beyond the range of the data. Regression analysis is an attempt to establish a relationship within a range. The relationship may, in fact, change radically over a different range.

e. Since in regression analysis we are minimizing the deviations about the mean, regression analysis is sensitive to changes in the mean. The mean may be distorted by errors in the sample data or by extreme points at the edge of the range of the data. Hence, erroneous results may occur.

14. For more information see: Walpole, Ronald E., et al. *Probability and Statistics for Engineers and Scientists,* 8th ed. Upper Saddle River, NJ: Prentice Hall, 2007.

SECTION SEVEN

DECISION ANALYSIS

(Return to Table of Contents)

1. Introduction.

a. To make a decision is to commit resources to a course of action. The ability to decide how to commit resources is surely one of the most important attributes of a commander or manager. Intuitively one wants to weigh the alternatives available in a decision situation to find the one that is best. This is usually not an easy task. Most important decisions are fraught with uncertainty, unknowns, and the unknowable. Key factors are imperfectly known. There may be uncertainty about the consequences of selecting a course of action. How does one weigh the alternatives? What is "best"? Decision analysis is an analytical technique for taking the risks, uncertainties, and decision maker's attitudes about values and risk into account. As the term is generally used, "decision analysis" implies that the decision maker is not confronted with a rational opponent, such as an opposing military leader or a commercial competitor, whose countermeasures to any action must be considered. If such opposition exists, the analyses move from decision analysis to the realm of game theory. In practice, this dichotomy is not absolute. In some circumstances, decision analysis techniques may have value even in the presence of a rational opponent.

b. The decision maker is central to decision analysis and his input will both inform and transcend the formal analytical techniques used. The analyst must recognize that the decision maker is powerfully influenced by his experience. He uses personal judgments, values, and quasi conscious reactions in making a decision. The unavoidable injection of the decision maker's personality into the decision circumstance can have both positive and negative effects. Positive effects, which are indispensable to a good decision analysis, include but are not limited to subjectively limiting the scope of an analysis, determinations of the relative importance of attributes inherent to the decision, and the inclusion of the decision maker's attitudes toward risk. Negative effects include very strong tendencies to make decisions based on mental rules of thumb that seem to operate below the level of human consciousness. There is substantial experimental evidence suggesting that these rules of thumb, more formally known as cognitive biases, powerfully effect everyone and cause people to make many decisions in a manner that is demonstrably irrational. Even the most objective decision maker may well be influenced by them. Decision analysis techniques should be applied in such a manner as to facilitate the positive input of the decision maker and to mollify the pernicious effects of subconscious bias. Since the personal input of the decision maker is so powerful, different decision makers, when placed in identical situations, may well make different decisions. This does not mean that one would be right and the other wrong. It could be that each of these people has made what is the "right" decision for him, but because of differing values and attitudes the "right" decision may differ.

c. Applying quantitative techniques in such a subjective area is simple in concept, but can be difficult in practice. Nevertheless, the process of decision analysis is extremely important, and can be a most useful management tool for the decision maker. In the discussion of decision analysis that follows, a logical framework for the decision process

which explicitly considers risk and risk preference, without trying to predict the decision maker's behavior, will be presented.

e. Decision analysis is intended to make it easier for the decision maker to make a good decision. Since this is the case, it is worthwhile to consider what a "good" decision is. It is quite tempting to define a good decision as one that produces a good result. On closer examination, this definition is inadequate. Have you ever encountered a circumstance where someone acted on what was clearly a "wild guess" and had things fall right into place out of pure luck? Did that person make a good decision, or was he just lucky? At the other extreme, some people make decisions after careful investigation and reflection, only to have things go wrong, again as a result of chance. Did they make bad decisions because the outcomes were bad? Judging the quality of a decision solely by the quality of the outcome is a little simplistic.

f. Perhaps the definition of a "good" decision should focus on what goes on during the decision process, rather than what happens after the decision is made. A good decision can be thought of as one where all of the pertinent information is gathered and considered in a systematic manner, within the time available. Decision analysis techniques are essentially tools that make it easier to identify the information that is needed, and evaluate it in a consistent, reasonable fashion. Their use does not guarantee a good outcome on any particular decision. Rather, through consistent application of such tools, a decision maker can expect to make decisions that result in good outcomes a higher proportion of the time than would be the case in the absence of decision analysis.

2. The Decision Environment

a. Decisions are made about the future. There is a choice between various alternatives. The objective is either to maximize or minimize some outcome or set of outcomes. In military analyses, the outcomes are often described as measures of effectiveness. Consider a private sector marketing decision, the objective might be to maximize profits, where the amount of profit achieved is considered the outcome or the measure of effectiveness. The objective might involve more than one measure of effectiveness. For instance, the company might want to maximize profits but also achieve a high degree of market penetration. For a military procurement decision, the objective might be to minimize cost, or it might be a minimization of cost and the attainment of a certain level of operational capability. After a decision is taken, things beyond the control of the decision maker take place which affect the outcomes arising from the alternative chosen. These uncontrollable occurrences are called states of nature. The state of nature determines the value of the outcome achieved by each alternative. In a financial investment example, suppose the alternatives (for simplicity's sake) are "buy stock" and "buy bonds". Which alternative is best depends on what happens to the stock market. If the stock market goes up significantly, a greater profit would probably be made if the "buy stock" alternative had been chosen. On the other hand, if the stock market goes down significantly, a greater profit would probably be made had the "buy bonds" alternative been chosen. The direction and magnitude of market movement are beyond the control of the decision maker and represent states of nature.

b. Decision analysis is classified into single and multi-attribute decisions.

(1) Single attribute decision analysis is used in circumstances where one attribute is of overwhelming importance and the decision can safely be made with reference to this attribute alone. The techniques used to analyze such decisions are dependent upon the level of knowledge one has about states of nature that may occur after the decision is made.

a. Condition of Certainty. Certainty is the condition which contains the most information about the states of nature and hence the most information about the consequences of the decision. This condition exists when the state of nature that will occur is known exactly and the value the measure of effectiveness will assume for each alternative is known or can be calculated. As a result, decisions made under the condition of certainty are usually either routine or trivial. If one knows with certainty exactly what the stock market is going to do, one would not need to be very smart to make the best decision. For this reason, no further reference to decision making under the condition of certainty will be made.

b. Condition of Uncertainty. When certainty is not present, levels of knowledge are sometimes categorized using ideas developed in 1921 by Frank Knight of the University of Chicago. Knightian uncertainty exists when no probability information is available regarding the possible states of nature. Suppose in the financial example, there are three possible states of nature: the market goes up 100 points this year; it remains the same; or it loses 100 points this year. Assume the profit that would be made under each state of nature can be calculated (a loss is simply a negative profit). If there is no information about the probabilities that the market will go up, stay the same, or go down, the investment decision will be made under the condition of uncertainty.

c. Condition of Risk. If the decision maker can assign usable probabilities to the possible states of nature associated with a decision, he or she is operating under a condition of Knightian risk. If outcomes associated with each course of action can be estimated, a type of weighted average outcome called an expected value can be calculated for each alternative.

(2) Multi-attribute decision analysis consists of the use of many attributes to evaluate alternatives. Many of the complex problems that the military encounters fall under this classification. The descriptions of attributes vary in units, scale, qualification, and importance. We will look at these later in our study.

3. Decision making Under Risk and Subjective Estimation.

a. In making decisions under the condition of risk, a Payoff Matrix like the one in Figure 1, will be constructed.

	Demand	
	High (0.8)	**Low (0.2)**
Small Complex	**8**	**7**
Medium Complex	**14**	**5**
Large Complex	**20**	**-9**

Figure 1

Suppose a development company has purchased land and must select one of three complex sizes (Small, Medium, Large) to build. In the selection process they must consider two demand levels (High, Low), with their associated probabilities of occurrence state in parenthesis. The above table summarizes the profit, in millions of dollars, for each complex/demand combination. Of course, a payoff matrix can enumerate more than three alternatives and two states of nature. This problem has been kept small for easy discussion.

b. When information which allows assignment of a probability of occurrence for each state of nature is available, the decision is said to be made under the condition of risk. Decision making under the condition of risk has two parts; assigning the probabilities of occurrence to the states of nature, and evaluating the information available to arrive at a choice of an alternative.

c. It is often the case when making decisions under the condition of risk that the probabilities of occurrence associated with the possible states of nature are known or can be calculated. In gambling situations, such as cards or dice, it is possible using classical probability theory and principles, to calculate the probability of drawing a full house, filling an inside straight, or rolling seven. In other situations, there may be enough empirical or experimental data available to calculate or at least approximate the probabilities using statistics. In these situations, the analyst's task is relatively easy. These situations are those for which most information is available upon which to base a decision. Nevertheless, there is risk involved since knowing the probabilities and having a favorable outcome occur is not the same thing.

d. How can probability estimates be obtained when the information available is severely limited? The answer is by making an educated guess. Many people feel that this is the weakest point in decision analysis.

e. In any decision situation, if hard facts or obvious probabilities are not available, it makes sense to get estimates of the probabilities in question from experts in the particular field of interest. Even when "objective data" are available, a strong argument can be advanced on behalf of using subjective expert opinion to weigh the implications of that information. Many people wish to discount subjective estimates simply because they are subjective. For many years it was commonly believed that any kind of quantitative analysis must be purely objective. Today it is realized that one cannot always separate the objective from the subjective. In addition, some people think that the effort to separate the two should not be made—at least not with the goal of saving the objective and discarding the subjective.

The use of subjectivity in a quantitative analysis is an attempt to get a handle on what is commonly called the "gut feeling" of an expert. It is an effort to use intangible as well as tangible information, to put feeling into a study, to utilize everything available to come to a proper conclusion.

f. In the example above, the probabilities given were derived by looking at historical data. The percent of times a high, moderate, and low demand were observed are used as the probability that they will again occur.

4. Decisions Under Risk: Expectation and Expected Value.

a. Once the probability of occurrence has been assigned to each state of nature (by whatever means), expected value is frequently used to select the best alternative when making decisions under conditions of risk because expected values allow both the probability of occurrence and the payoffs (or consequences) if the event does occur to be taken into account. The expected value of any alternative is the sum of the products of the payoffs and the probabilities associated with those outcomes.

b. Returning to the contract example, the normal analysis, or normal form, is to use a payoff matrix as in Figure 2, below.

	Demand		
	High (0.8)	**Low (0.2)**	**Expected Value**
Small Complex	**8**	**7**	**7.8**
Medium Complex	**14**	**5**	**12.2**
Large Complex	**20**	**-9**	**14.2**

Figure 2

Now that probabilities of demand level have been estimated, the expected values of the alternatives have to be computed. This is done by multiplying the respective payoffs by the probabilities and summing across.

For Small Complex EV = (8)(0.8) + (7)(0.2) = 7.8

For Medium Complex EV = (714)(0.8) + (5)(0.2) = 12.2

For Large Complex EV = (20)(0.8) + (-9)(0.2) = 14.2

Since this is a profit table, the objective is to maximize the profit associated with the complex, when using expected value as the basis for selection the decision maker would select the alternative with the highest expected value, the **Large complex alternative**.

c. Theoretically, the payoffs for an alternative associated with the possible states of nature, together with the probabilities of occurrence constitute a discrete probability distribution. In statistics, the expected value and the arithmetic mean are the same thing. The same considerations which apply when using the mean to make decisions based on a population probability distribution are appropriate when considering whether or not expected

value is appropriate for analyzing a particular decision.

d. First, expected value is a statistical average. If repeated decisions of the same type are made, the average outcome will be better if based on expected value than will be the case if any other decision criterion is used. However, no other criterion will make the right decision as frequently as will expected value and when those other criteria are in error, those errors will tend to be more costly than will those made by using expected value. Therefore, when a large number of similar decisions must be made expected value as the decision criterion will be a better solution than any other decision criterion chosen. This assumes of course that the decision maker is reasonably adept at estimating the probabilities of occurrence. The key here is that a large number of decisions must be made which gives the outcomes the opportunity to "average out".

e. Many students find the preceding paragraphs disappointing. They had hoped for a technique that guarantees a good decision. Decision analysis does not pretend to provide such a technique. Its purpose is to organize relevant information; identify significant relationships; help identify those elements which are critical to the decision, and maintain a consistency of choices throughout the decision process. There are, however, no guarantees.

f. Expected Value Using Excel. Figure 3 presents a Single Attribute Template file name "DA SingleAttrib Template.xlt" that is available for download in the ORSA Net Website.

	A	B	C	D	E	F	G	H	I
1			Single Attribute Decision Analysis						
3	Input Data in Yellow Fields								
4	Input Number of Alternatives (MAX = 6)					NOTE: Regrets Are in Dark Blue.			
5	Input Number of States of Nature (MAX = 6)								
6	Input Type of Problem (High or Low)								
7	Input Analysis (1=Uncertainty 2=Risk)								
8									
9		Enter Prob.							
10									E(VALUE)
11		State/ Alternative							
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									
24									
25			Best Value	Choice					
26			0.00	Alt. 1					
27									
28		MINIMAX REGRET =	0.00						

Figure 3

Inputs are placed in the yellow areas of the template. First place the number of alternatives and states of nature in the first two fields (E4, E5). You can place up to 6 of each in this template. Next select the type of problem you have (E6). Type "High" if the data you have are benefits, in other words the higher the number the better. Type "Low" if the data are costs, or the lower the number the better. Next place a 1 under input analysis (E7) if you are under a condition of uncertainty, you don't have probabilities, and a 2 if you are under condition of risk, where you have known probabilities for the states of nature. In this case, probabilities are available, so use 2. If you are in a state of risk enter the probabilities for each state of nature, otherwise leave blank (C9 thru H9, as necessary). In the green areas, you can type in the name of each alternative (B12, B14, B16, etc, down to B22, as necessary) and state of nature (C11 thru H11, as necessary). Finally enter the data within the table (C12 thru H12, C14 thru H14 etc, down to C22 thru H22, as necessary).

The completed template appears in Figure 4 below.

	A	B	C	D	E	F	G	H	I
1				Single Attribute Decision Analysis					
3	Input Data in Yellow Fields								
4	Input Number of Alternatives (MAX = 6)				3	NOTE: Regrets Are in Dark Blue.			
5	Input Number of States of Nature (MAX = 6)				2				
6	Input Type of Problem (High or Low)				High				
7	Input Analysis (1=Uncertainty 2=Risk)				2				
8									
9		Enter Prob.	0.8	0.2					
10			State 1	State 2					E(VALUE)
11		State/ Alternative	High	Low					
12	Alt. 1	Small	8	7					7.8
13			12.0000	0.0000					
14	Alt. 2	Medium	14	5					12.2
15			6.0000	2.0000					
16	Alt. 3	Large	20	-9					14.2
17			0.0000	16.0000					
18									
19									
20									
21									
22									
23									
24									
25			Best Value	Choice					
26		E(VALUE)	14.20	Alt. 3					
27									
28									

Figure 4

The blue figures on rows 13, 15, and 17 are regrets. They do not pertain to this problem and will be addressed later. They may be ignored at this juncture. The template has calculated the expected values which appear in column I. With a bit of rounding, these expected values match the values calculated just below Figure 2. The small table in the range B25 to D28 reports the best expected value and the alternative selected.

6. Decisions Under Uncertainty:

a. In the previous section, it was assumed that probabilities were available describing the likelihood of each state of nature. In many situations, usable probabilities may not be available. For example, when purchasing a house, most people would not know the numerical probabilities that the house would appreciate or depreciate by any given amount. Most people could not attach a meaningful probability to the rise or fall of a mutual fund or an investment in gold or silver. When deciding on the appropriate strategy, usable numerical probabilities might be quite difficult or impossible to obtain. For example, when the United States made the decision to invade Iraq in the aftermath of the 9/11 attacks, the desired outcome was the prompt establishment of a stable, democratic government, but it is doubtful that a numerical probability could be assigned to this event. Decisions made under these circumstances lend themselves to examination under Knightian uncertainty. Three of the more common techniques for decision analysis under uncertainty will be discussed. These techniques can be thought of alternately as techniques to inform a decision to be made in the future in the absence of probability information, or ways to understand the innate inclination of the decision maker. As was mentioned above, decision analysis cannot be divorced from the personal experience and inclinations of the decision maker. This inextricable association of the technique with the intuitions and feelings of the decision maker is apparent in the techniques that follow.

b. Optimistic Approach (Also called maximax when dealing with benefits, or minimin when dealing with costs). Some decision makers are innately optimistic. They view the world and the decisions they confront as instances of opportunity, where much can be achieved. They are not inclined to take the counsel of their fears. People with this mind set are likely to focus on the most favorable outcomes that present themselves in a decision circumstance. This is particularly true in the absence of firm probability information suggesting that the best outcome may not be the most likely to occur, or may even be quite unlikely. This focus on the best payoff tends to lead such decision makers toward the alternative that leads to that payoff. The two step approach formalizes this decision inclination. Figure 5, below, presents the same payoff information about the complex problem that appeared in Figure 1, above. Here, we are assuming that the decision maker does not have useable probabilities. The first step in an optimistic (or maximax in this case) selection is to pick the best payoff associated with each alternative. Since this payoff matrix presents profits, the best payoff is the highest number for each alternative, or 8, 14, and 20 for Small, Medium, and Large, respectively. The second step is select the best outcome from these three payoffs. The best outcome is "20", which suggests a Large complex is the best.

	Demand		Optimistic	Conservative
Alternative	High	Low	Best	Worst
Small	8	7	8	7
Medium	14	5	14	5
Large	20	-9	20	-9

Figure 5

c. Conservative Approach (also called maximin when dealing with benefits, or minimax when dealing with costs). Other decision makers can be thought of almost as psychological opposites of the optimistic personalities described above. They are cautious, wary individuals who are quite concerned in avoiding the really bad outcomes that might follow from a decision. The two step conservative approach formalizes this mindset. The decision maker reviews each payoff associated with an alternative and selects the worst. Referring to Figure 5, this yields 7, 5, and -9, respectively. Step two involves comparing these numbers and selecting the "best of the worst". In this instance, 7 is the best and suggests the "Small" alternative as the selection that would most match the inclinations of this decision maker.

d. Minimax Regret Approach. This approach focuses on minimizing how badly a decision maker feels after he has selected an alternative that results in something less than the best possible payoff. It begins with converting a payoff matrix into a regrets matrix. In this procedure, the decision maker selects the best payoff that could result from each state of nature. In Figure 5, where the payoffs are profits, these are 20, and 7 for high, and low demand, respectively. He then finds the difference between the best payoff for each state of nature and the actual payoff for that state of nature that would result from the selection of each alternative.

For example, if the choice was "Large" and high demand occurred, the difference between the actual payoff and the best payoff possible would be 20 – 20 = 0. There would be no regret, as the selection of "Large" yielded the best possible result associated with the state of nature that actually happened. If the decision maker had chosen "Medium", the regret would be 2- 14 = 6. If he had chosen "Large" instead of "medium", he could have profited by an additional 6 million dollars. The 6 is taken as a measure of how badly the decision maker feels after the fact. Following this procedure throughout Figure 5 yields the regret matrix shown in Figure 6, below.

	Demand		MiniMax Regret
Alternative	High	Low	Worst
Small	20 – 8 = 12	7 – 7 = 0	12
Medium	20 – 14 = 6	7 – 5 = 2	6
Large	20 – 20 = 0	7 – -9 =16	16

Figure 6

Once the regrets are determined, the decision maker selects the highest regret level for each alternative. The highest regret for "Small" is 12; for "Medium", 6; and for "Large", 16. He then selects the alternative that minimizes the "maximum regret". In this case, the "Medium alternative does so.

e. Some people feel these techniques are so simplistic that they either would not be applicable to real world decision scenarios or would have little to offer. While understandable, this is not a universally held opinion. There is considerable commentary in the literature that suggests that a very large number of decisions, perhaps a majority, are made in circumstances where useful probability information is not available. For example, decisions

faced by entrepreneurs frequently involve judgments about activities that have never been attempted before. There is no statistical or experiential basis for assignment of probabilities. A military analog to such activity. is the development of new weapons systems. Hiring decisions involve an estimation of the likelihood that the applicant will be successful. In many cases, neither the applicant nor the employer can attach meaningful numeric probabilities to this decision. Strategic and tactical decisions involve the assessment of likelihood of success but are likely to be made in the absence of probabilities. The techniques discussed above could present useful methods for ranking alternatives in situations like these.

f. In addition to their uses as decision tool, these approaches could be thought of as ways to describe behavior patterns either exhibited by individuals or desired in certain circumstances. Successful military commanders are frequently credited with being quick to commit to bold action when circumstances called for audacity. Douglas MacArthur's invasion of Inchon during the Korean war was certainly audacious. MacArthur chose to invade at a location where logistical support would be severely constrained by extreme tides, but had the potential of placing a large combat force far to the rear of the North Korean army. This decision focused on the considerable "payoff" of success and could be viewed as optimistic or maximax behavior.

The financial crisis of 2007 – 2008 can be viewed through this lens. Virtually all banks held large investments in securities whose value depended on the payment of mortgages by homeowners. Since World War II, default rates had been very low and there had never been an instance of the entire housing market loosing value at the same time. As a result, the banking industry attached very low probabilities to the occurrence of a significant increase in defaults. For a variety of reasons, underwriting standards and the types of mortgages being written had changed so dramatically that the actual probability of significant defaults was much higher than the industry imagined. When housing prices began to fall, default rates increased dramatically and the value of the mortgage backed securities held by the banks plunged. The industry transitioned from Knightian risk to uncertainty as it became apparent nobody knew what the actual probability of mortgage defaults was. This meant nobody knew the probability of a bank failing due to their exposure to such securities. Overnight interbank lending virtually ceased as no bank could determine the probability that another bank would fail and not repay the loan. Banks hoarded capital and nonbank companies with sterling credit could not get loans at reasonable rates. Companies with lesser credit ratings could not get loans at all. This industry wide refusal to lend is classic conservative maximin behavior caused by the absence of probability information. Although it may have made sense for individual banks, it was very destructive to the national economy.

These examples are not meant to imply that optimistic behavior is universally preferable to conservative. Indeed, Custer could be thought of as an optimistic, maximax decision maker and the technique did not serve him well in his last battle. A conservative, maximin orientation on the part of an aircraft design team is likely a good idea. Which techniques should be used is dependent on the inclinations of the decision maker, and the circumstances of the decision.

g. Decisions under Uncertainty Using Excel: Figure 7 presents the Single Attribute Template. This is the same template as appears in Figure 3. The numbers in Figure 7 are from the scenario presented in Paragraph 4, a. (above). To use the template for single

attribute analysis under uncertainty, enter the number of alternatives, "3" in this case, in cell E4, the number of states of nature "2", in cell E5, and indicate that high numbers are good by placing a "High" in E6. In E7, place a "1" indicating the decision is being made under uncertainty. Enter the states of nature descriptions in cells C11 through D11 and the alternative descriptions in cells B12, B14, and B16. Place the payoffs to the right of the alternative descriptions in cells C12, and D12 for the "Small" alternative and in the corresponding cells for "medium" and "large". Note that probabilities are not required. The output of the template is displayed in the small table in cells B25 through D28. An optimistic decision maker would pick the lowest payoffs by alternative, (20, 14, 8), then select "20" as the best of the best. This is the "Best Value" reported in cell C26. Alternative 3 is the choice indicated in cell D26. A Conservative decision maker would select the worst payoffs by alternative, (7, 5, -9), then pick the best of the best, or "7". This is the "Best Value" reported in cell C27. Alternative 1 is indicated as the selection.

	A	B	C	D	E	F	G	H	I
1	**Single Attribute Decision Analysis**								
3	Input Data in Yellow Fields								
4	Input Number of Alternatives (MAX = 6)				3		NOTE: Regrets Are in Dark Blue.		
5	Input Number of States of Nature (MAX = 6)				2				
6	Input Type of Problem (High or Low)				High				
7	Input Analysis (1=Uncertainty 2=Risk)				1				
8									
9		Enter Prob.							
10			State 1	State 2					
11		State/ Alternative	High	Low					
12	Alt. 1	Small	8	7					
13			12.0000	0.0000					
14	Alt. 2	Medium	14	5					
15			6.0000	2.0000					
16	Alt. 3	Large	20	-9					
17			0.0000	16.0000					
18									
19									
20									
21									
22									
23									
24									
25			Best Value	Choice					
26		OPTIMISTIC-MAXIMAX =	20.00	Alt. 3					
27		CONSERVATIVE-MAXIMIN =	7.00	Alt. 1					
28		MINIMAX REGRET =	6.00	Alt. 2					

Figure 7

The blue numbers in cells C13 through D13, C15 through D15, and C17 through D17 are regrets. They were calculated using the procedure illustrated in Figure 7 (above). Once the regrets are available, the decision maker would pick the highest regret for each alternative, or 12, 6, and 16 in this case. He would then "minimize the maximum regrets" by selecting the lowest regret from these three. In this case, the minimum regret is 6, associated with Alternative 2. The minimum regret and the alternative selected are reported in cells C28 and D28.

7. Decision Trees.

a. Although the payoff matrix introduced previously can be useful for analyzing small decision problems, it becomes very cumbersome for larger problems. The decision tree provides a method for analyzing decisions sequentially by providing a pictorial representation of the decision problem. Such a technique is termed extensive in that the multistage nature of the decision is represented explicitly.

b. Structurally, there are three fundamental elements to a decision tree:

(1) **Decision node,** □ . A square is used to represent a decision point. The decision maker must make a choice between alternative courses of action at a decision node.

(2) **Chance node,** ◯. A circle represents nature's choice. At a chance node, one of any number of possible states of nature takes place. In effect, nature decides which state of nature occurs.

(3) **Branch, "-------"**. A straight line is used to connect decision nodes and chance nodes. The branch is used to describe the choice. A branch leaving a decision node represents an alternative and a branch leaving a chance node represents a chance occurrence.

c. Figure 8, illustrates the complex example in decision tree form. The six outcomes listed in the payoff matrix are simply paths along the branches from left to right. The payoffs are listed at the tips of the branches in the tree.

d. There are three steps to follow in constructing a decision tree. First, the analyst lays out all possible scenarios using the appropriate nodes and branches. In the case of the complex example, the entire scenario is short. Initially, a decision will be made as to which complex to select. After that decision is made, the complex is built and demands on the system made.

e. After the scenario is laid out, the second step is to determine the probabilities associated with the states of nature that arise at each chance node. These probabilities are generally written on the branch which with they are associated. The probabilities of a high, or low demand, 0.8 and 0.2 respectively, appear on the branches starting at the three chance nodes.

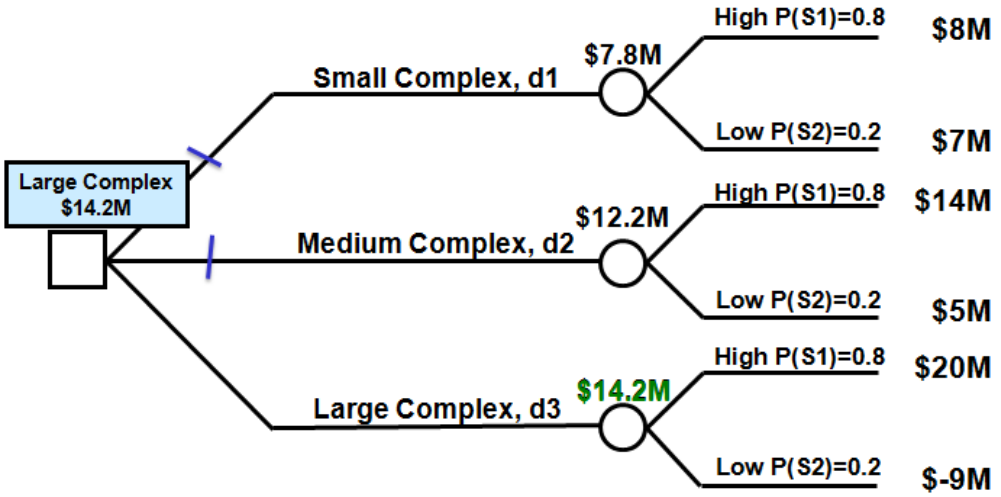


Figure 8

f. The final step is to determine the costs and payoffs associated with each of the endpoints on the right hand extreme of the tree. A second "trip through the tree" can be helpful here, as the various branches represent things that either have to be done or have happened. Usually, the analyst will have some idea of the costs and payoffs associated with these things, and a review of the tree can result in a list of information that has to be collected. In Figure 8, the numbers at the far right extreme of the decision tree are the payoffs associated with each final outcome (expressed in profit).

g. The decision problem is solved by beginning at the tip of the last branches on the right and working to the left using a technique sometimes referred to as "folding back." When folding back a decision tree, different actions are taken at the two different kinds of nodes. In Figure 8, the first nodes encountered in the right to left folding back process are chance nodes. When a chance node is encountered, the expected value of the chance occurrence, the node represents, is calculated. In this case, the expected value of the upper chance node is:

For Small Complex EV = (8)(0.8) + (7)(0.2) = 7.8

The expected value for the second chance node is:

For Medium Complex EV = (714)(0.8) + (5)(0.2) = 12.2

The expected value for the lower chance node is:

For Large Complex EV = (20)(0.8) + (-9)(0.2) = 14.2

The expected values of the chance nodes are assigned to those nodes and are used for later operations in the folding back process. In the Figure 8, the expected values have been written above the chance nodes they pertain to.

h. Moving further to the left across the tree, the next node encountered is a decision node. Here a conscious decision is required. The technique presumes the decision maker is

standing at this point in the scenario, looking to the right in the tree and assessing his options. Based on expected value, the best decision is the Large complex with an expected value of $14.2 million.

i. Dr Paul Jensen of the University of Texas has created among others (see appendix B on add-ins for download information) an add-in for decision trees. The following is the same example as above using his add-in. Once you activate the add-ins, select New Problem from the decision analysis menu. An input dialog like figure 9 will appear. Make the proper inputs as shown in the figure; the number of nodes includes the decision node, all chance nodes, and the terminal nodes.

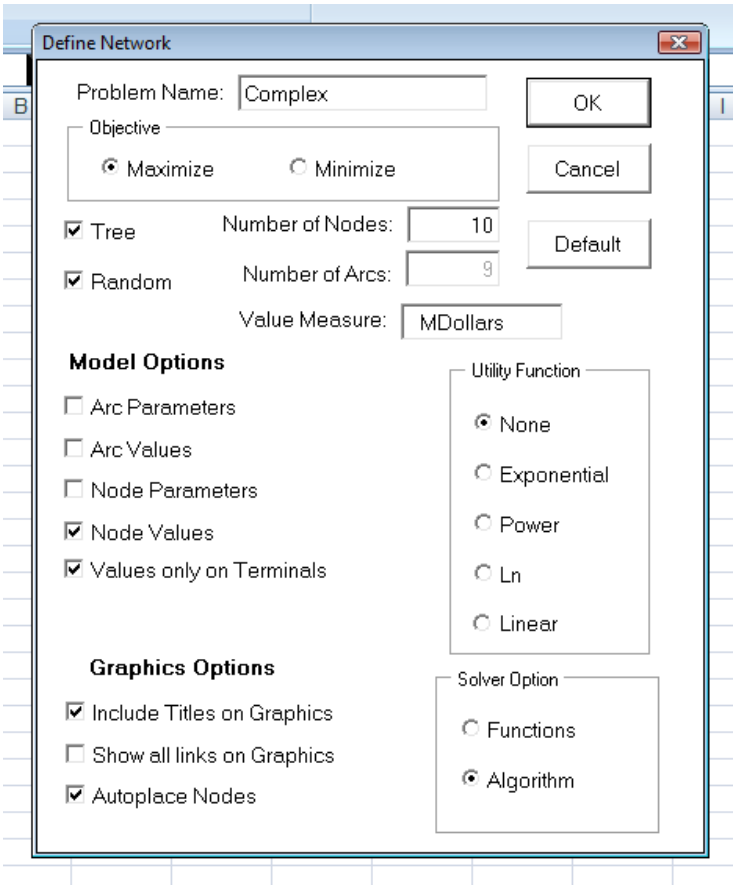


Figure 9

Once you click "ok" a template is created as in figure 10. The template will be filled in; you will need to make the changes to fit your problem. Each node is numbered; you need to identify it as a decision, chance or terminal node in the right set of cells as well as the value at each terminal node. In the left set of cells identify how the nodes 'fit' together, and the probabilities.

Decision Analysis Model

Change Structure
Change Node/Arc
Solve
Graphics
Sort Network

Name:	Complex	Titles:	Yes	Arc Params:	No	Utility:	None
Type:	DA	All Links:	No	Arc Values:	No	Tree:	No
Solver:	Algorithm	Autoplace:	Yes	Node Params:	No		
Goal:	Max			Node Values:	Yes		
				Term. Values:	Yes		

Arc Data

Arc Index	From Index	To Index	Arc Prob.	From Name	To Name	Arc Name	Arc Total
1	1	2	0	Complex	Small	Small	0
2	1	3	0	Complex	Medium	Medium	0
3	1	4	1	Complex	Large	Large	14.2
4	2	5	0.8	Small	High	High	6.4
5	2	6	0.2	Small	Low	Low	1.4
6	3	7	0.8	Medium	High	High	11.2
7	3	8	0.2	Medium	Low	Low	1
8	4	9	0.8	Large	High	High	16
9	4	10	0.2	Large	Low	Low	-1.8

Node Data

Node Index	Node Name	Node Type	Node $Dollars	Optimum Value	Optimum Arc	Optimum Decision	Node Level	Node Depth
1	Complex	D	0	14.2	3	Large	0	0
2	Small	C	0	7.8			1	0
3	Medium	C	0	12.2			1	1
4	Large	C	0	14.2			1	2
5	High	T	8	8			2	0
6	Low	T	7	7			2	1
7	High	T	14	14			2	2
8	Low	T	5	5			2	3
9	High	T	20	20			2	4
10	Low	T	-9	-9			2	5

Figure 10

Once the template is filled in, click on the red button "solve". On the right set of cells you will see the values associated with each node under optimum value; for each chance node these are the expected values. For each decision node you will see the decision made; in this case Med (for Medium Contract). If you click on the red button "graphics", the tree will be drawn in a new tab as shown in figure 11.

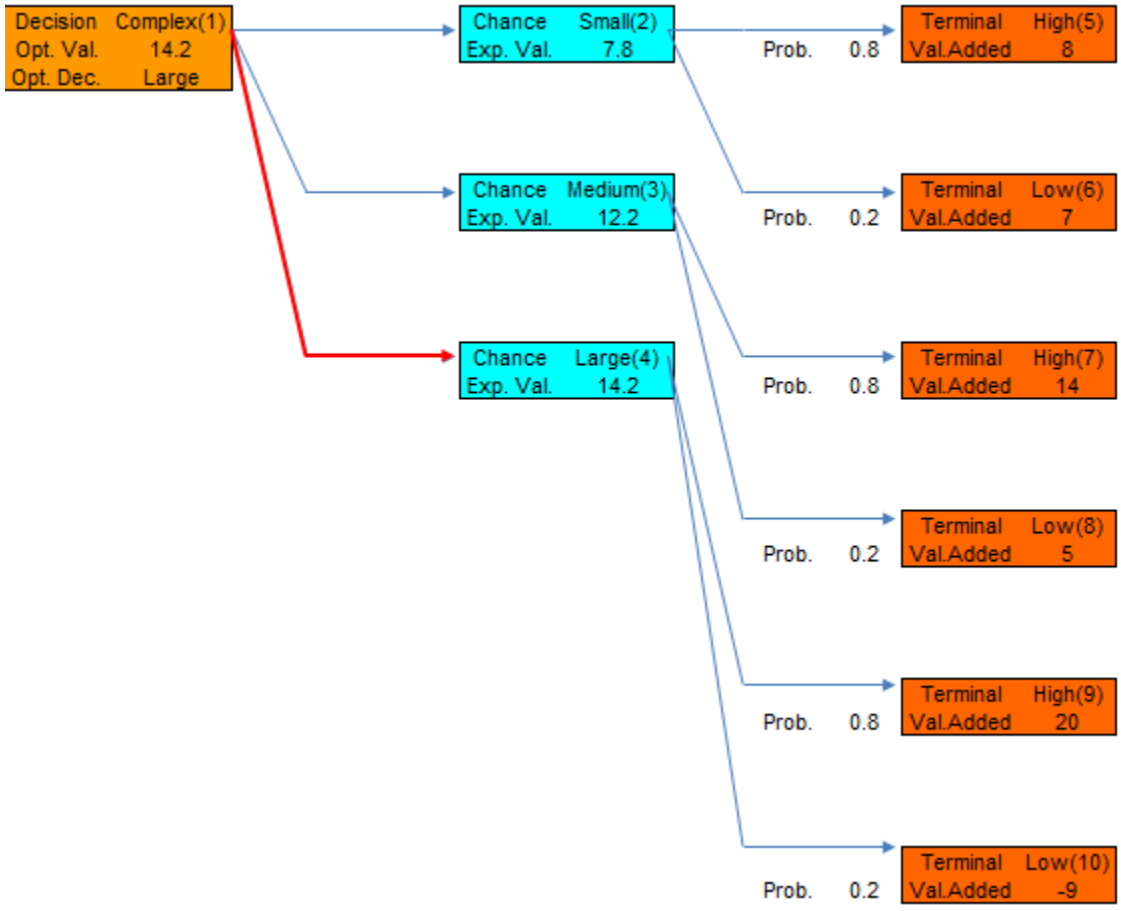


Figure 11

Although there are more sophisticated programs and add-ins for building decision trees, they are not free. This add-in works well enough for relatively small problems and it is free to use.

j. Some advantages to using a decision tree verses tables are:

- It is less difficult to represent different probabilities for the same states of nature for each alternative.

- You can easily represent different states of nature for the alternatives.

- You can incorporate decision and chance nodes at multiple levels throughout the tree.

k. The decision tree can serve as an excellent organizer in a decision problem. The thought process required to draw the tree forces one to think through possible future scenarios in a more systematic manner than might otherwise be the case. This type of review causes the analyst to become more thoroughly familiar with the intricacies of the problem than he would be, had he not attempted the technique. Once drawn, the tree demonstrates the probability, cost, and payoff information that is required to make the decision, and thus provides guidance in the information gathering effort. Finally, with very little explanation from the analyst, the decision tree can be understood by managers outside the analytical community, and can become a useful visual communications tool in some instances

7. Sensitivity Analysis.

a. Typically the data you collect is not perfect and therefore requires additional examination. Sensitivity analysis allows you to examine inputs you may be wary about and determine if changes in those inputs would impact the final decision. In Single attribute decision analysis, those changes can be made to payoffs and well as probabilities. The basic steps are as follows:

- Identify values that are uncertain or you think may change
- Estimate Extreme values (Highest and Lowest possible values)
- Recalculate Expected Value for each alternative at each extreme
- Check if decision changes at each value
- If decision does not change, then value is not sensitive
- If decision does change, find value at which you are indifferent (where expected value is the same for the alternatives)
- Determine if the change point is important (compare to original value)

b. There are several ways to conduct a sensitivity analysis depending on your familiarity with algebra and/or analysis programs. We will do a sensitivity analysis on the probabilities in the complex problem in three ways; a table, algebra, and a graph.

c. Table. We already said we'd perform a sensitivity analysis on the probabilities, since probabilities range from 0 to 1, those are our extreme points. First we'll evaluate the expected values for each alternative for each extreme.

P(High)	P(Low)	EV(Small)	EV(Medium)	EV(Large)
0	1	7	5	-9
.8	.2	7.8	12.2	14.2
1	0	8	14	20

Figure 12

Recall that the initial probabilities were 0.8 and 0.2 for a high and low demand respectively. If we change the probability of high to 1, then the probability of low would be 0 (probabilities across all states of nature must add to 1). The expected values are shown in Figure 12 and the Large complex would still be selected. However, changing the probability of high to 0 and low to 1, note that the Small complex would be selected. This tells us that somewhere between a probability of 0 and 0.8 for the high demand, the decision changes. We need to seek that probability. We can select a probability in between, say 0.5, and evaluate again. Figure 13 shows that at 0.5 the Medium complex would be selected, so in fact there are two change points.

P(High)	P(Low)	EV(Small)	EV(Medium)	EV(Large)
0	1	7	5	-9
.5	.5	7.5	9.5	5.5
.8	.2	7.8	12.2	14.2
1	0	8	14	20

Figure 13

If we now evaluate a probability of 0.25 for high as in figure 14, we see that the expected values for the Small and Medium complexes are the same. This is our first change point.

P(High)	P(Low)	EV(Small)	EV(Medium)	EV(Large)
0	1	7	5	-9
.25	.75	7.25	7.25	-1.75
.5	.5	7.5	9.5	5.5
.8	.2	7.8	12.2	14.2
1	0	8	14	20

Figure 14

Evaluating the probability of high at 0.75 still yields Large as the selection as shown in Figure 15. Going to a probability of 0.7 shows the expected value for the Medium and Large complexes to be the same. This is our second change point.

P(High)	P(Low)	EV(Small)	EV(Medium)	EV(Large)
0	1	7	5	-9
.25	.75	7.25	7.25	-1.75
.5	.5	7.5	9.5	5.5
.7	.3	7.7	11.3	11.3
.75	.25	7.75	11.75	12.75
.8	.2	7.8	12.2	14.2
1	0	8	14	20

Figure 15

There for we can say if the probability of a high demand:

- is between 0 and 0.25, select the small complex.
- is between 0.25 and 0.7, select the medium complex.
- is between 0.7 and 1, select the large complex.

Lastly, we compare this to the original probability. It is highly unlikely that if you did some good analysis you'd be as far off as a probability of 0.25. Therefore we can be assured that we would not select the small complex. However, given all the circumstances of your analysis, could a probability of 0.7 occur? This is a question for you and the decision maker to answer. How comfortable are you selecting a large complex at a 0.8 probability if a decision change occurs at a 0.7 probability?

d. Algebra. Let's look at how to derive the change points using an algebraic method. All the steps remain the same except the method used to gain the probabilities.

Let p = P(High), then:
P(Low) = 1 – P(High) = 1-p

We can therefore express the expected values as follows:

EV(small) = P(High)(8) + P(Low)(7) = P(High) + [1-P(High)](7) = 8p + 7(1-p) = p+7
EV(medium) = 14p + 5(1-p) = 9p + 5
EV (large) = 20p + (-9)(1-p) = 29p – 9

We can now solve for p by equaling equations to each other:

EV(small) = EV(medium)
p+7 = 9p+5
8p = 2
p = 0.25

EV(medium) = EV(large)
9p+5 = 29p-9
20p = 14
p = 0.7

Note we get the same answer as the table above.

e. Graph. The last method is to estimate the probabilities is by drawing a graph. You can use excel or graph paper to plot the graph as in Figure 16.

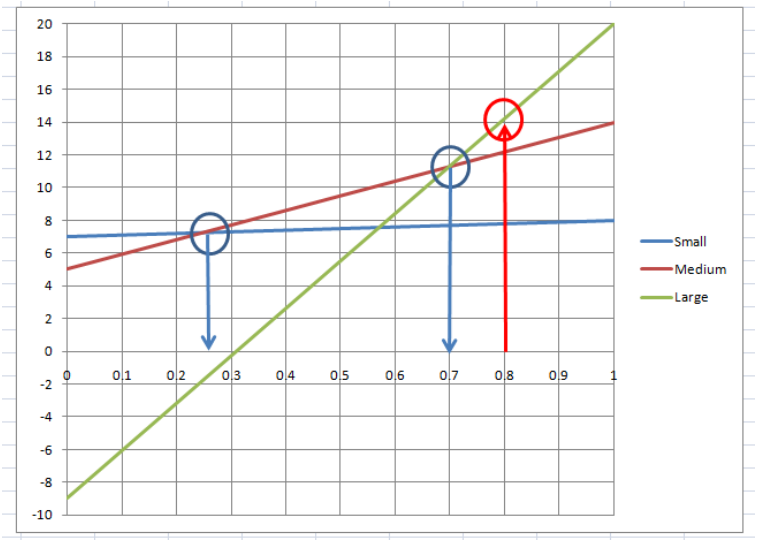


Figure 16

Since this is a plot of profits we are interested in the high expected values and follow the lines of small until it intersects with medium at a probability of 0.25, then continue following the medium line until it intersects large at 0.7. We also plotted the original probability of 0.8. A graph is very useful to show the change points to the decision maker in a very simple and understandable fashion.

8. Multi-attribute Decision Analysis.

Many problems involve more than one attribute under different states of nature. In fact, many of the problems we face have many attributes with different criterion that are evaluated for each alternative.

a. Characteristics of multi-attribute decision analysis:

(1) Presence of qualitative data. One of the first things we may have to deal with is the type of data collected. Although we would prefer to have quantitative data, the fact is that there are times where only qualitative data is available for some attributes. This type of data usually deals with qualifying attributes as poor, good, excellent, etc. Obviously, this type of data is non numerical and mathematical operations cannot be performed.

(2) Different units of measure. Next, decision problems typically involve many characteristics described using different units of measure. Units such as dollars, gallons, miles, pounds, speed, heights, and such cannot be meaningfully added or subtracted. Comparison of quantities expressed in different units of measure in a decision context can be challenging. For instance, the selection of an automobile with high horsepower probably also implies higher gasoline costs. Does the increased power compensate for the increased costs? Answering such questions can be difficult when considering just two characteristics and many decision problems involve multiple characteristics.

(3) Scales. Along with the different types of units, we encounter scaling issues. If we have for example a number like $1,000,000 as an attribute and another like 180 miles, the much larger number will over power the smaller making it insignificant in the analysis, Therefore, numbers must be scaled in a proportional manner.

(4) Conflicting attributes. Many times we may have to deal with attributes that conflict. As an example we want to minimize cost, so smaller numbers are better, yet we want to maximize cargo, so larger numbers are better. Again, we must adjust these numbers so that an increased value always represents desirability.

(5) Preferences. Lastly, not all attributes are of equal importance to the decision maker. It is necessary to determine which attribute is deemed most important, then determine the importance of "lesser" attributes in comparison.

b. We will continue the discussion of multi-attribute decision analysis by using the example in Figure 17. Suppose we have selected the following six systems as alternatives in a new helicopter program. We have also decided upon the attributes that we will use to evaluate the alternatives and collected the appropriate data.

	Cruise Speed	Climb Rate	Payload	Reli-ability	Maneuver-ability	Maintain-ability	Cost
A1	**145**	**580**	**4625**	**High**	**High**	**10**	**3.0**
A2	**175**	**415**	**2750**	**Avg.**	**Avg.**	**12**	**4.9**
A3	**190**	**500**	**2700**	**High**	**Low**	**9**	**2.5**
A4	**150**	**450**	**2550**	**Very High**	**High**	**6**	**2.5**
A5	**140**	**425**	**2500**	**Avg.**	**Avg.**	**14**	**5.1**
A6	**130**	**400**	**2510**	**Avg.**	**Avg.**	**7**	**2.4**
	Ben	**Ben**	**Ben**	**Ben**	**Ben**	**Cost**	

Figure 17

c. Screening Techniques. Before beginning any mathematical manipulation of attributes, we can look at two screening techniques. These techniques allow us to eliminate alternatives in advance.

(1) Satisficing. An alternative meets satisficing criteria if the alternative meets minimum requirements for **all** attributes where a minimum requirement is specified.

(2) Dominance. An alternative dominates another if its attribute values beat or tie those of the other alternative. *REMEMBER: Larger Benefits are Better - but so are Smaller Costs!*

(3) Applying the screening techniques.

a. Satisficing. Suppose now that there is a requirement that the cruise speed of the helicopter must be at least 135 knots. Looking under that attribute, note that the cruise speed of system A6 is 130. Because it does not meet a specified requirement, the system is eliminated from the analysis.

b. Dominance. Dominance is applied by doing pairwise comparisons of each system versus the others. First look at system A1 vs. A2. Note that A1 is lower in cruise speed but higher in climb rate, therefore there is no dominance between those systems. We continue this process with each alternative. Evaluating A1 vs. A5 we note that A1 is better across all criteria, system A5 can therefore be eliminated from the analysis; thus yielding Figure 18, below.

	Cruise Speed	Climb Rate	Payload	Reli-ability	Maneuver-ability	Maintain-ability	Cost
A1	**145**	**580**	**4625**	**High**	**High**	**10**	**3.0**
A2	**175**	**415**	**2750**	**Avg.**	**Avg.**	**12**	**4.9**
A3	**190**	**500**	**2700**	**High**	**Low**	**9**	**2.5**
A4	**150**	**450**	**2550**	**Very High**	**High**	**6**	**2.5**
	Ben	**Ben**	**Ben**	**Ben**	**Ben**	**Cost**	

Figure 18

d. Evaluating Cost. Cost of a system is not a Measure of Effectiveness (MOE). It is not a performance measure of the system. Although is it an important criterion, we do not want to include it at this point in the analysis. First we want to evaluate the systems based on performance. We will then perform a cost benefit analysis to determine performance vs. cost (bang for the buck).

e. Qualitative to quantitative value transformation. The next step is to transform the qualitative word data into numerical form. We will use the chart in Figure 19 (top of next page). This table is being used for the purpose of this instruction. You are not limited to this scale. You may opt to use a scale from 1 to 100 for example. A larger scale will give more separation amongst the values. In addition you can get the decision maker involved in determining the scale. There is no requirement that the spread between each value be equal.

Qualitative Value	Score
Very High	5
High	4
Average	3
Low	2
Very Low	1

Figure 19

After applying our scale, our table now looks like this:

	Cruise Speed	Climb Rate	Payload	Reli-ability	Maneuver-ability	Maintain-ability
A1	**145**	**580**	**4625**	**4**	**4**	**10**
A2	**175**	**415**	**2750**	**3**	**3**	**12**
A3	**190**	**500**	**2700**	**4**	**2**	**9**
A4	**150**	**450**	**2550**	**5**	**4**	**6**
	Ben	**Ben**	**Ben**	**Ben**	**Ben**	**Cost**

Figure 20

e. Proportional scaling. Even though all our data is numerical now, we still have units of measure and scaling issues. To illustrate units of measure difficulties, operational cost is measured in dollars, range in miles, and reliability, commonality, and mobility in dimensionless scores. There is no way to add this data together without modification of the information. With respect to scaling issues, consider reliability and operational cost. The magnitude of the cost numbers is so large relative to the reliability scores, that the cost influences would overwhelm the reliability influences in subsequent calculations unless further adjustments are made. We will use a technique called simple additive weighting. This technique deals with both of these issues. It also transforms all attributes into quantities where larger numbers are desirable. We use the following formulas to scale our data. If the attribute is one where a larger value is desired, scale using this formula:

$$\text{Rescaled Score} = \frac{\text{Attribute Value}}{\text{Best Attribute Value}}$$

If a larger value is desired, then the "Best Attribute Value" would be the highest value of that particular attribute across all the alternatives. The "Attribute Value" would be the value of the attribute for each of the alternatives. The calculation would be repeated for all of the values of the attribute being scored across all the alternatives.

If the attribute is one where a smaller value is desirable, scale using this formula:

$$\text{Rescaled Score} = \frac{\text{Best Attribute Value}}{\text{Attribute Value}}$$

If a small value is desired, then the "Best Attribute Value" would be the lowest value of that particular attribute across all the alternatives. The "Attribute Value" would be the value of the attribute for each of the alternatives. This calculation would also be repeated for all of the values of the attribute being scored across all the alternatives.

To transform the value "cruise speed", first note that this attribute is a Maximize or benefit attribute. A higher value is better, so the "Best Attribute Value" is 190. We will take each alternative's attribute value and divide it by this value.

$$\text{A1} = \frac{145}{190} = 0.76 \qquad \text{A2} = \frac{175}{190} = 0.92 \qquad \text{A3} = \frac{190}{190} = 1.00 \qquad \text{A4} = \frac{150}{190} = 0.79$$

Now let's look at "maintainability". A lower value is desirable. The "Best Attribute Value" is 6. We will take that value and divide it by each alternative's attribute value.

$$\text{A1} = \frac{6}{10} = 0.60 \qquad \text{A2} = \frac{6}{12} = 0.50 \qquad \text{A3} = \frac{6}{9} = 0.67 \qquad \text{A4} = \frac{6}{6} = 1.00$$

Note that in each case the best value always receives a score of 1; all others will be proportionally scaled between 0 and 1.

Scaled information is presented in Figure 21, below.

	Cruise Speed	Climb Rate	Payload	Reli-ability	Maneuver-ability	Maintain-ability
A1	**0.76**	**1.00**	**1.00**	**0.80**	**1.00**	**0.60**
A2	**0.92**	**0.72**	**0.59**	**0.60**	**0.75**	**0.50**
A3	**1.00**	**0.86**	**0.58**	**0.80**	**0.50**	**0.67**
A4	**0.79**	**0.78**	**0.55**	**1.00**	**1.00**	**1.00**

Figure 21

f. Attribute Importance. At this point we could add the values across the table and recommend an alternative. However, this would assume that all the values are of equal importance. Because, we usually consider some attributes more important than others, we apply weights to the different attributes. One method is to question either the decision maker or some panel of experts to seek the weights. First, we would have them list the attributes in order of importance or rank them. Then the most important attribute is assigned a score of 100. We then ask them to assign a value to the second attribute based on the first being 100. If the second attribute is half as important as the first, it receives a score of 50. If the third attribute is 25% as important as the most important attribute, it receives a score of 25, and so on. They continue scoring in this manner until the last attribute is scored. Once done, the scores are added, and then each score divided by the total to give the appropriate weight. Let's assume, after applying this process, we came up with the weights in Figure 22:

Attribute	Rank	Raw Score	Weight
Cruise Speed	6	35	35/420 = 0.08
Climb Rate	4	65	65/420 = 0.15
Payload	1	100	100/420 = 0.24
Reliability	2	90	90/420 = 0.21
Maneuverability	3	70	70/420 = 0.17
Maintainability	5	60	60/420 = 0.14
Total		420	

Figure 22

After adding the decimal weights, the evolving table might look like Figure 23.

	Cruise Speed	Climb Rate	Payload	Reli-ability	Maneuver-ability	Maintain-ability
Weight	0.08	0.15	0.24	0.21	0.17	0.14
A1	0.76	1.00	1.00	0.80	1.00	0.60
A2	0.92	0.72	0.59	0.60	0.75	0.50
A3	1.00	0.86	0.58	0.80	0.50	0.67
A4	0.79	0.78	0.55	1.00	1.00	1.00

Figure 23

The next step is to apply the weights. We will multiply each of the scaled attributes by their appropriate weight and then we add across to find the final score. For example, the final score for Alternative A1 that appears in the rightmost column of the Figure 24 was determined in this manner:

(.08 x .76) + (.15 x 1.00) + (.24 x 1.00) + (.21 x .80) + (.17 x 1.00) + (.14 x .60) = .880

Final scores for Alternatives A2, A3 and A4 were calculated in the same manner.

	Cruise Speed	Climb Rate	Payload	Reli-ability	Maneuver-ability	Maintain-ability	Weighted Score
Weigh	0.08	0.15	0.24	0.21	0.17	0.14	
A1	0.064	0.155	0.238	0.171	0.167	0.086	0.880
A2	0.077	0.111	0.142	0.129	0.125	0.071	0.654
A3	0.083	0.133	0.139	0.171	0.083	0.095	0.706
A4	0.066	0.120	0.131	0.214	0.167	0.143	0.840

Figure 24

g. Alternative selection. We can now make a selection. By using this method, the alternative with the highest score is always selected. In this case the highest score is 0.880, therefore, alternative A1 is selected. Note however, that some scores are relatively close. To make a good selection, the selected value should be 5 to 10 percent higher than the second score. In this case, A1 is only 4% higher than A4. Remember that the technique is permeated with subjectivity. There is subjectivity with the qualitative data, there may be some estimates in the quantitative data, and the weights are also subjective. In a case like this, you may need to do sensitivity analysis on the weights, and values to assess the amount of confidence you can attach to the results.

h. Applying cost. If you recall, we had cost data for each alternative. The scores above reflective a performance score of the helicopters. We can now take those scores and compare them to the cost of the helicopters (Figure 25).

	Cruise Speed	Climb Rate	Payload	Reli-ability	Maneuver-ability	Maintain-ability	Weighted Score	Cost
Weigh	0.08	0.15	0.24	0.21	0.17	0.14		
A1	0.064	0.155	0.238	0.171	0.167	0.086	0.880	3.0
A2	0.077	0.111	0.142	0.129	0.125	0.071	0.654	4.9
A3	0.083	0.133	0.139	0.171	0.083	0.095	0.706	2.5
A4	0.066	0.120	0.131	0.214	0.167	0.143	0.840	2.5

Figure 25

We can graph the scores vs. the cost as in Figure 26.

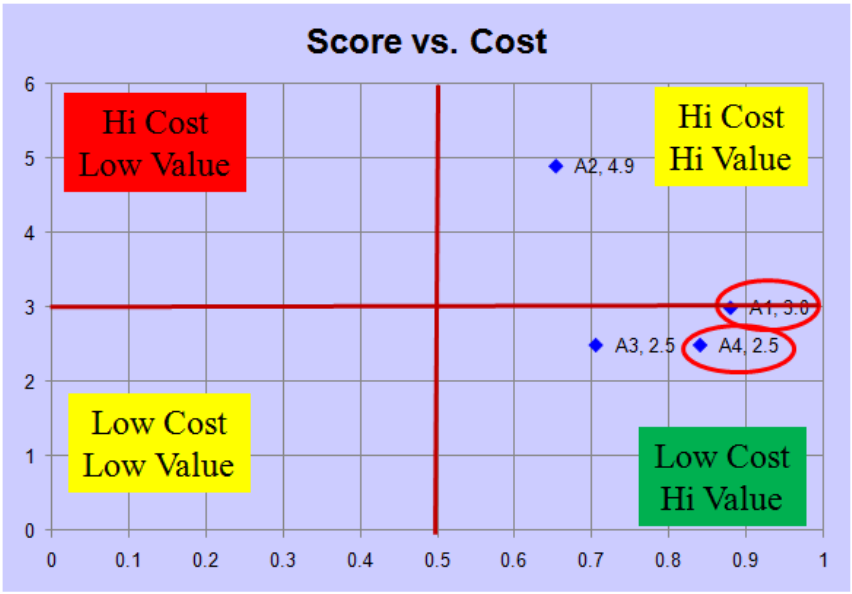


Figure 26

Dividing the graph in quadrants assists with the analysis. Best case is to have high valued systems at relatively low cost. We already said that helicopters A1 and A4 are not very different in terms of performance value, however, A4 cost 0.5 million dollars less. The question then becomes, is the performance gain of A1 worth the additional cost? If we look back at the data, the biggest contributor to the value of A1 is payload. It is nearly double the amount of the other systems. You can then ask the question, is having twice as much payload worth 0.5 million dollars? That has to be answered by the decision makers and operators and depends upon the operational environment and use of the system.

9. Multi-Attribute Decision Analysis in Excel.

The template that appears in Figure 27 (above) facilitates simple additive weighting for multi-attribute decision analysis. It can be downloaded from ORSA Net and is called "DA, SAW Template.xlt". This template allows for 8 alternatives and criteria (attributes). Place the number of alternatives and attributes in cells E5 and E6, respectively. Attribute titles can be placed in cells D11 through K11. Alternative titles are placed in cells C14 through C21. In cells D11 through K11, you must place a 1, if the criteria is a benefit (that is higher values are better), and a 2 if the criteria is a cost (that is lower values are better). This allows the program to use the appropriate conversion formula. Type-in the values for each alternative, under each attribute, in cells D14 through K21, as necessary. You must convert any qualitative data into numbers before entering them into the template.

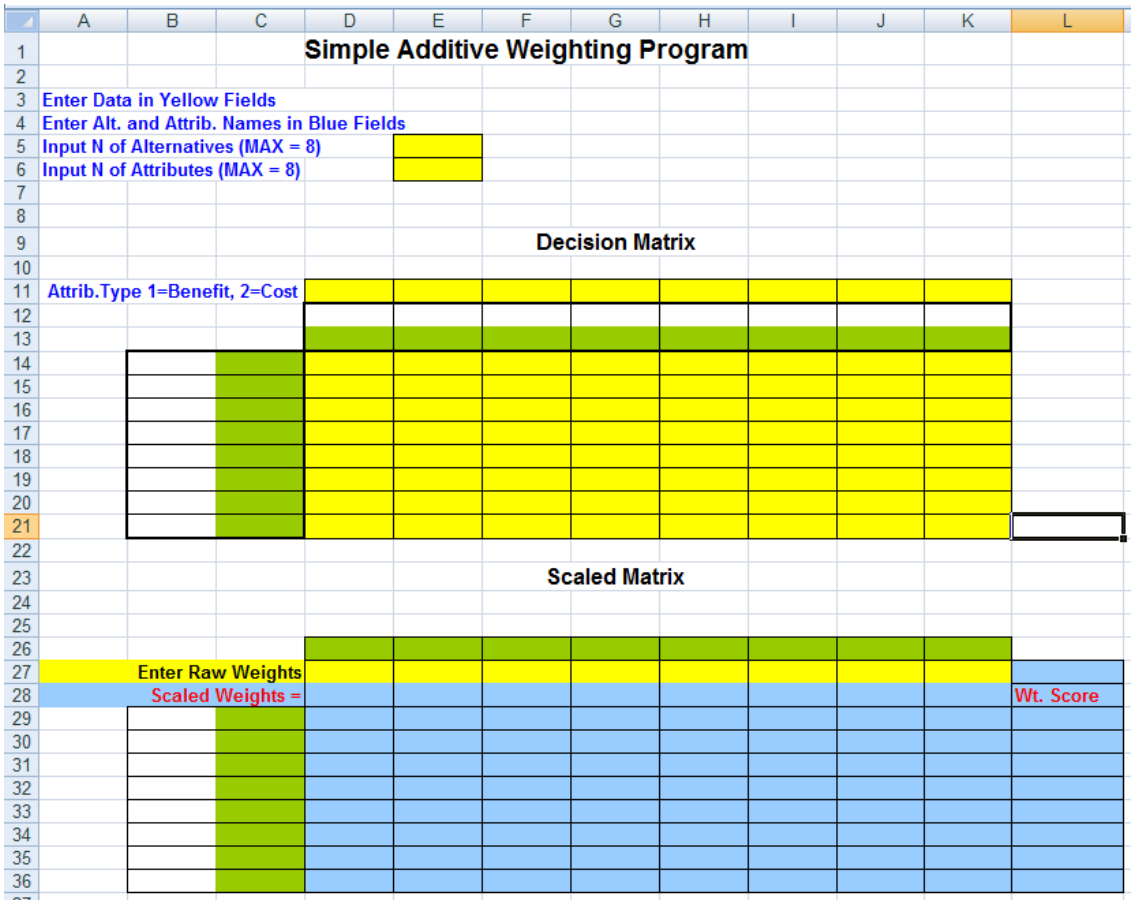


Figure 27

The "Decision Matrix" portion of the template, with the input values, appears in Figure 28.

Simple Additive Weighting Program

Enter Data in Yellow Fields
Enter Alt. and Attrib. Names in Blue Fields
Input N of Alternatives (MAX = 8) 4
Input N of Attributes (MAX = 8) 6

Decision Matrix

Attrib.Type 1=Benefit, 2=Cost		1	1	1	1	1	2	
		Attrib. 1	Attrib. 2	Attrib. 3	Attrib. 4	Attrib. 5	Attrib. 6	
		Speed	Climb	Load	Reliable	Maneuver	Maintain	
Alt. 1	A1	145	580	4625	4	4	10	
Alt. 2	A2	175	415	2750	3	3	12	
Alt. 3	A3	190	500	2700	4	2	9	
Alt. 4	A4	150	450	2550	5	4	6	

Figure 28

Figure 29 shows the "Scaled Matrix" portion of the template (which is located below the "Decision Matrix" portion). The decision makers responses to questions concerning the relative importance of the attributes are entered below each attribute name on row 27. The template converts these responses into decimal weights and calculates the overall score for each alternative. The difference between theses scores and those in Figure 25 are due to rounding.

Scaled Matrix

		Attrib. 1	Attrib. 2	Attrib. 3	Attrib. 4	Attrib. 5	Attrib. 6			
		Speed	Climb	Load	Reliable	Maneuver	Maintain			
Enter Raw Weights		35	65	100	90	70	60			
Scaled Weights =		0.08	0.15	0.24	0.21	0.17	0.14		Wt. Score	
Alt. 1	A1	0.7632	1.0000	1.0000	0.8000	1.0000	0.6000		0.8803	SELECT
Alt. 2	A2	0.9211	0.7155	0.5946	0.6000	0.7500	0.5000		0.6541	
Alt. 3	A3	1.0000	0.8621	0.5838	0.8000	0.5000	0.6667		0.7057	
Alt. 4	A4	0.7895	0.7759	0.5514	1.0000	1.0000	1.0000		0.8409	

Figure 29

Figure 30 presents the graphical display of the results that appears on the "Graph of Results" worksheet in the template.

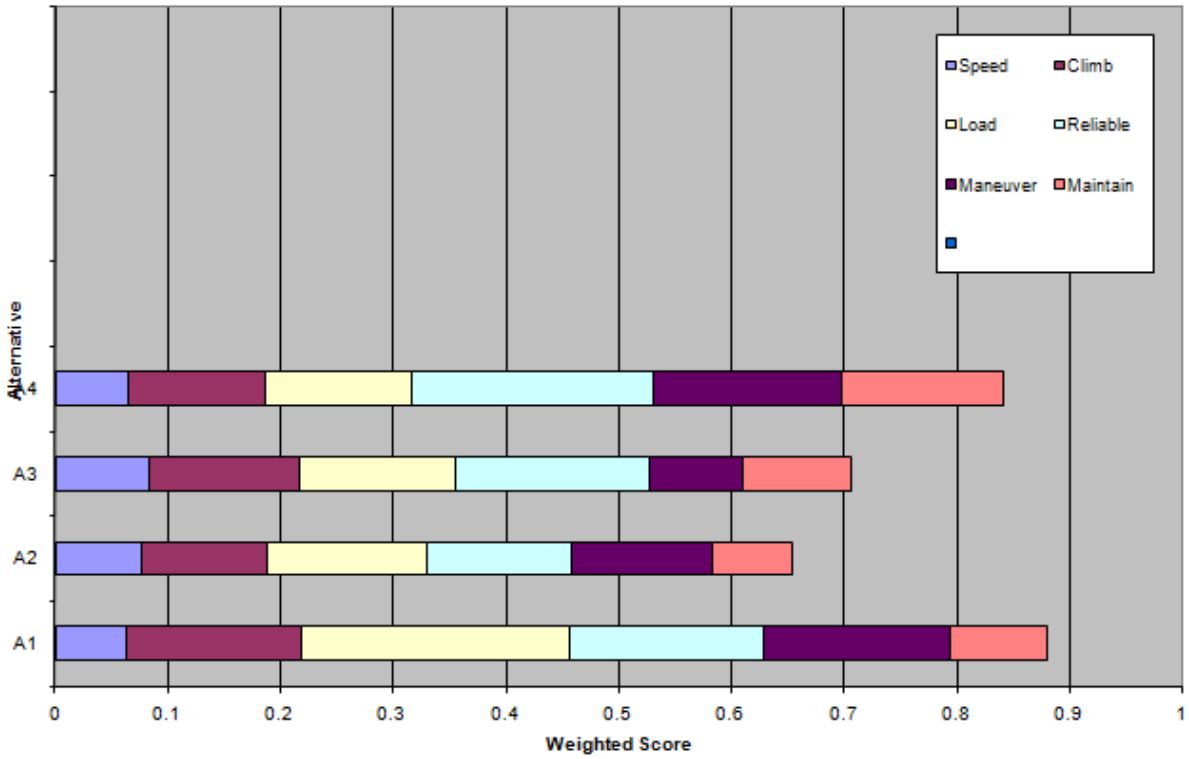


Figure 30

Note that you are given a bar graph that not only shows which alternative has the highest aggregate score, but the contribution of each criterion to the total score.

10. Value Focus Thinking.

Although we are not going into this topic, it is important to mention. It would be of great interest to study and as appropriate apply the concepts of value focus thinking into a formal decision analysis. The concept was created by Ralph Keeney in his book, "Value-focused thinking: a path to creative decision making". Two of the most beneficial concepts of the process are; first, the focus on the objective rather than the alternatives. By keeping focus on the objective, you can avoid bias and search for alternatives that best meet those objectives. The second concept is the use of utility functions. Instead of converting values as we did above, utility (or value) functions are created for each criterion. The values are then converted using these functions. This technique allows for how much a decision maker values each criterion based on the objectives rather than just a straight conversion. For example, if we take the payload criterion from the above example, although A1 has twice as much payload, he may not give that much value to the payload. He may say that anything above 3000 lbs does not provide any more value thereby bringing the other alternatives closer. We encourage that you continue studying decision analysis beyond what is written so you can provide the decision maker with sound, valuable, analytical advice and recommendation.

SECTION EIGHT
PROJECT MANAGEMENT
(Return to Table of Contents)

1. Introduction.

a. Project management is the discipline of planning, organizing, securing, and managing resources to achieve specific goals. A project is a series of tasks or activities required to accomplish an objective. Project management involves; Defining the project, Planning the activities, Scheduling the activities, Budgeting for funds, Controlling the project and Replanning and rescheduling as required. In this section will focus primarily on planning the activities by using either The Program Evaluation and Review Technique (PERT) or The Critical Path Method (CPM).

b . The Program Evaluation and Review Technique (PERT) is a method of scheduling, budgeting, and controlling resources to accomplish a predetermined job on or ahead of schedule. PERT also highlights both favorable and unfavorable developments in your project, allowing managers the opportunity to shift resources as early as possible when necessary.

c. PERT started in 1957 when it was used as an aid for controlling the progress of the Polaris ballistic missile program for the Navy. The Polaris program involved the management and control of a large number of diversified activities, including controlling and coordinating 250 prime contractors and over 9,000 subcontractors. As a result of using PERT as a planning, communication, control, and reporting tool, the Polaris project was completed over two years ahead of schedule.

d. The Critical Path Method (CPM) is also a method of scheduling and controlling resources. It was developed independently of PERT but concurrently. Both methods are interested in start and completion times of projects, tradeoffs between cost and time, and the interrelationships between jobs. However, CPM assumes that cost and time relationships are known with certainty while PERT allows for uncertainty. Because of these differences, PERT is used more in research and development projects where uncertainty must be analyzed through statistical techniques, while CPM is typically used in projects where some experience exists and the uncertainties have been reduced to negligible terms.

2. Construction of the PERT/CPM network.

a. Regardless of the method used, several characteristics apply.

(1) The project is usually a one-time or first-time project consisting of numerous well-defined activities or subtasks. When each of these individual activities or tasks is completed, the entire project is completed.

(2) Individual activities may be started or stopped independently of each other unless the start of one is dependent upon the completion of another. The beginnings and completions of activities are termed Events, and refer to a stage of completion of the project. Each event must be separately identifiable.

(3) An event merely marks the start or end of activities; it consumes no time or resources. It is a point in time. An activity (such as pouring the foundation for a building) does consume time and resources.

b. In order to apply PERT/CPM analysis to a given project the following steps are required:

(1) From management, or through use of subject matter experts and available data, obtain a verified table for the project that includes at least:

a. Name of Each Activity

b. Best Time Estimate for Each Activity

c. Immediate Predecessor(s) for Each Activity

(2) Use that table to create a network model:

a. Start at the left with a start node.

b. Draw from it an arc towards the right connecting each activity that has NO immediate predecessors.

c. Continue adding activities from their immediate predecessors connecting with arcs. Each activity node is rectangular in shape with the following representation:

Activity	Earliest Start	Earliest Finish
Activity Time	Latest Start	Latest Finish

d. Finally, "collapse" all activities that are NOT immediate predecessors into a single Finish Node on the right side of the network.

e. Work back through the network from right to left, checking that EACH activity is used—but only once—and that ALL immediate-predecessor relationships are accurately displayed.

f. Ideally, re-verify the table and network with management—and update both as necessary.

g. Using the FORWARD PASS procedure, calculate the earliest possible time that each activity could occur. The first activities that have no predecessors start at time 0. Each follow-on activity begins at the latest finish time of its predecessor. The activity time is added to the early start times to determine the earliest finish time for each activity. Activities that have more than one predecessor start with the greatest activity finish time. At the end of the forward pass you will have determined the expected completion time for the project.

h. Using the BACKWARD PASS procedure, calculate the latest possible time each activity could occur. All activities that were collapsed to the Finish node will start the backward pass with the end project time. Subtract the activity time to determine the latest start time. Use the latest start time as the latest finish time for predecessors. Predecessors with more than one activity following, are given the smallest of the latest times.

(3) Find the CP ("critical path," longest time path) for the network:

a. Subtract the latest start times from the earliest start times for each activity to determine the slack time.

b. Compare the overall path times, and choose the LONGEST time path as the network's CP. This path will be the one with all its activities having zero slack time.

c. If costs are available for each activity, add ALL of those costs to estimate overall project cost.

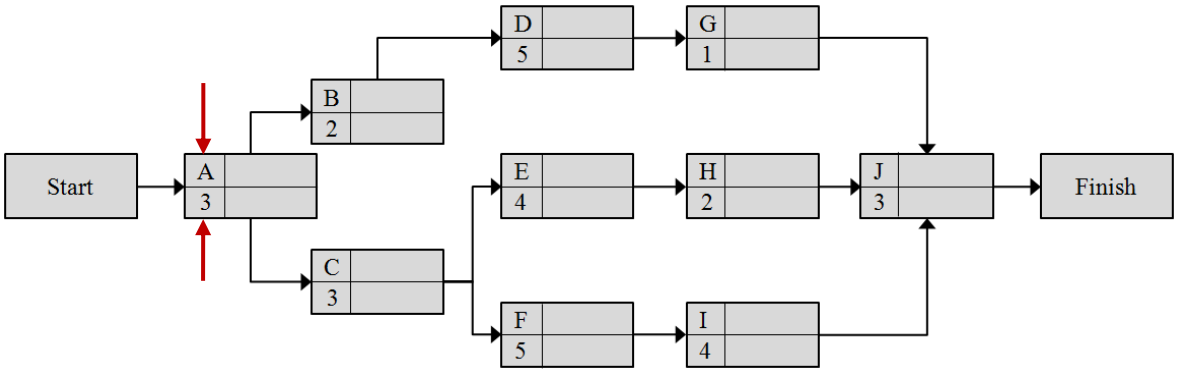
3. CPM Example.

a. Motor transport is the predominant mode of transportation for the reception, onward movement, and sustainment of forces. Motor transport units must be highly trained, rapidly deployable, and decidedly capable (FM 55-30, Chapter 1). Planning is key; however, time can be limiting factor in preparing. What activities are required in the planning process? How much time will planning require? Is there a method to better that time? What activities are critical to obtaining the shortest planning time possible? To answer these questions, let's construct an activity table, a network and then evaluate each activity accordingly.

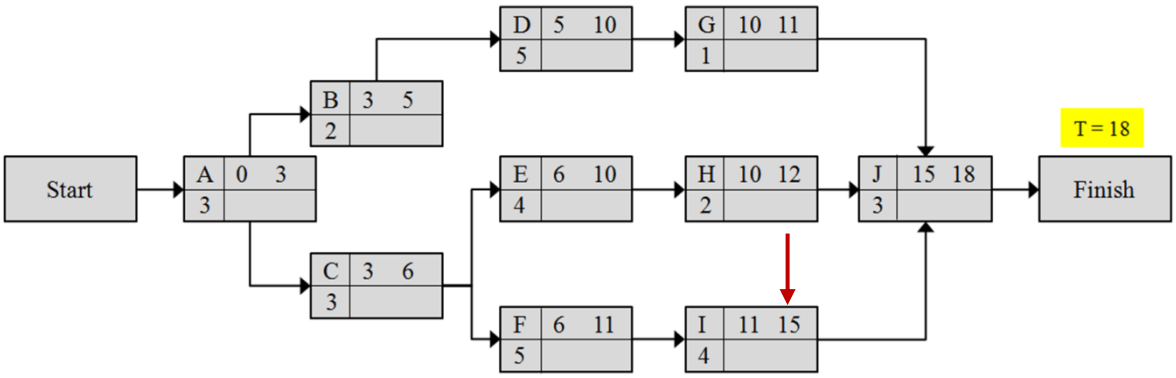
b. Construct an activity table (note that the activities are coded in the table; ideally you would create all your tasks and code them for simplicity of the diagram, i.e. A=Develop Concept of Movement). Completion time for this task, if all tasks are done sequentially, is 32 hours.

Task	Activity	Predecessor(s)	Time (Hrs)
Develop Concept of Movement	A	None	3
Conduct Risk Assessment	B	A	2
Prepare Vehicles	C	A	3
Validate Risk Assessment	D	B	5
Conduct Training - Active Defense	E	C	4
Conduct Training - Passive Defense	F	C	5
Publish Risk Assessment	G	D	1
After Action Review - Active Defense	H	E	2
After Action Review – Passive Defense	I	F	4
Conduct Final Check/Briefing	J	G, H, I	3
Total Time			32

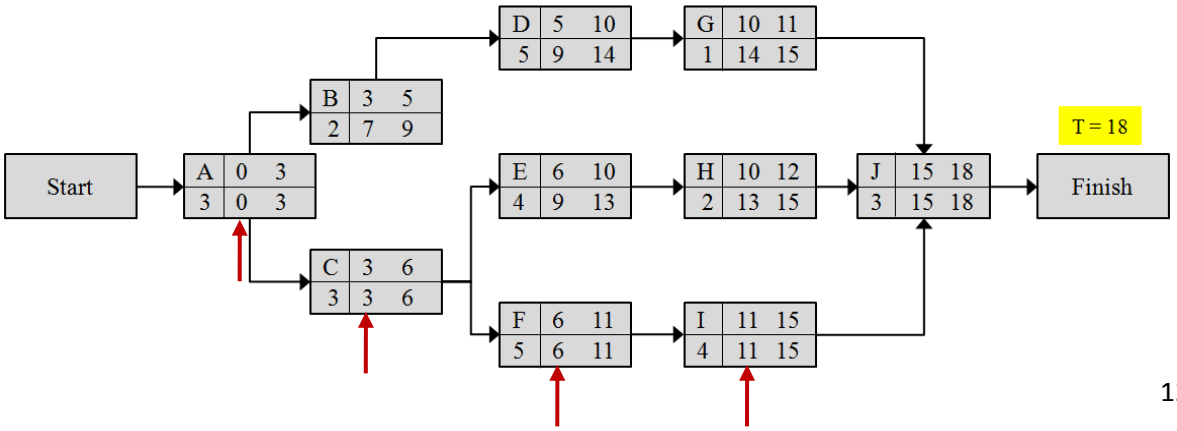
c. Develop the network. Draw a start node. Only activity A has no predecessors. Then activities B and C both follow A. Continue drawing in this manner. Be sure to also designate the Activity Time for each node as you complete the network.



d. Calculate the project's Finish Time by using the forward pass method. Start activity A at time 0, and then add its time of 3 to get the early Finish Time. Activities B and C can then both start at time 3, when A finishes. Continue this process forward. Note that activity J's Start Time is the greater of the Finish Times of G, H, and I.



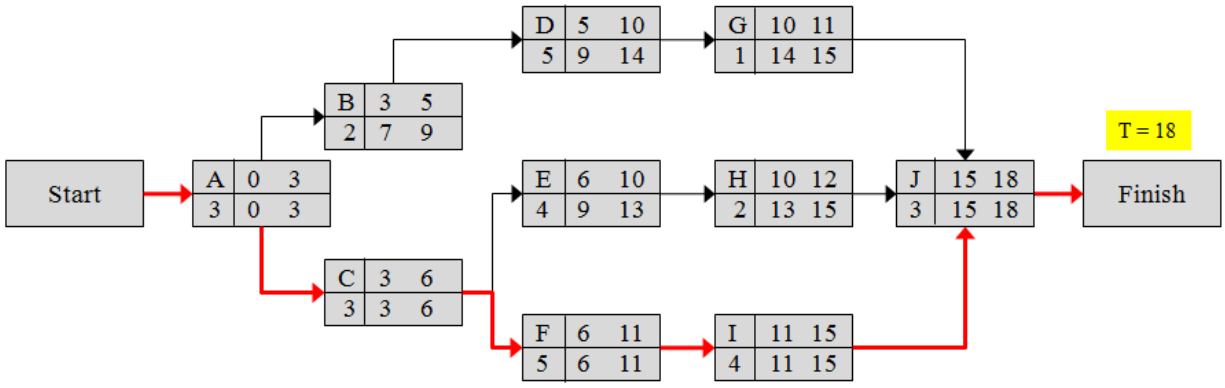
e. Conduct a backward pass and calculate Slack Times. Starting with the last Activity, J, give it the Latest Finish equal to the project end time, 18 hours (a substantial improvement over 32 hours). Then subtract the Activity Time to get the latest start time. Carry the Latest Start time to the Latest Finish time of activities preceding. Note than when an activity precedes more than one, the smaller of the two times is taken back, as in Activity C, where the 6 from activity F is taken back.



The Slack Times are then computed by subtracting the Early Start time from the Latest Start times. The following charts show Slack Times.

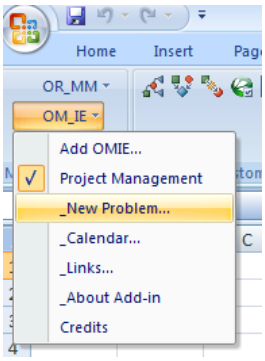
Activity	Slack Time
A	0 – 0 = 0
B	7 – 3 = 4
C	3 – 3 = 0
D	9 – 5 = 4
E	9 – 6 = 3
F	6 – 6 = 0
G	14 – 10 = 4
H	13 – 10 = 3
I	11 – 11 = 0
J	15 – 15 = 0

f. Determine the critical path. The activities along the critical path have no slack time. These are the activities that must be managed closely because any delay will cause the project to be delayed. Also note the project completion time is determine by the critical path. Additionally, if you are not satisfied with the end time, you can divert resources or add resources to activities along the critical path to shorten their times. Be aware that you must recalculate times after doing so, and the critical path may switch to one of the other paths.

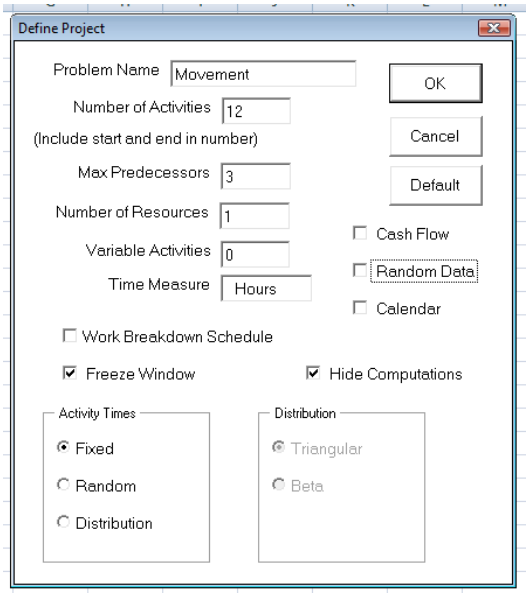


Critical Path = A, C, F, I, J

g. We can use Dr. Jensen's Excel add-in for project management (see app. B).



First, in the add-in tab, under the OM_IE add-ins ensure Project Management is activated and select New Problem.

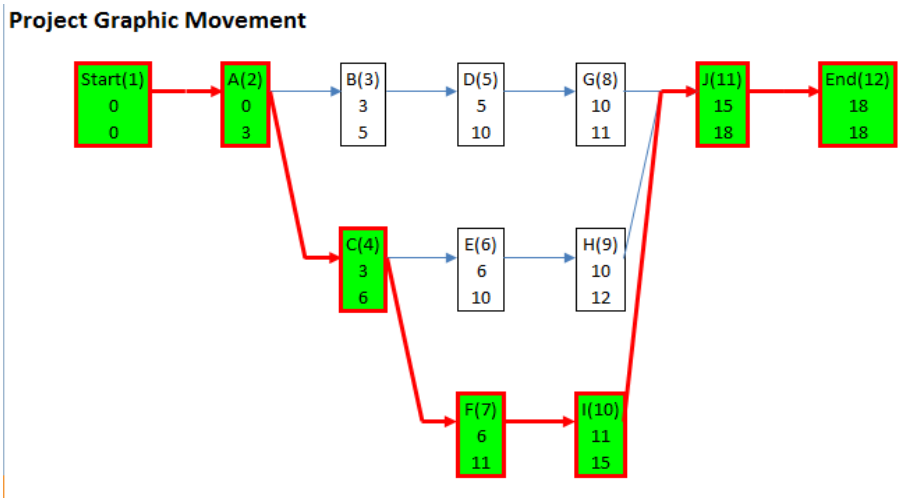


In the pop up window, make the appropriate inputs for the project, specifically the number of activities, max predecessors, and time measure. Then click ok.

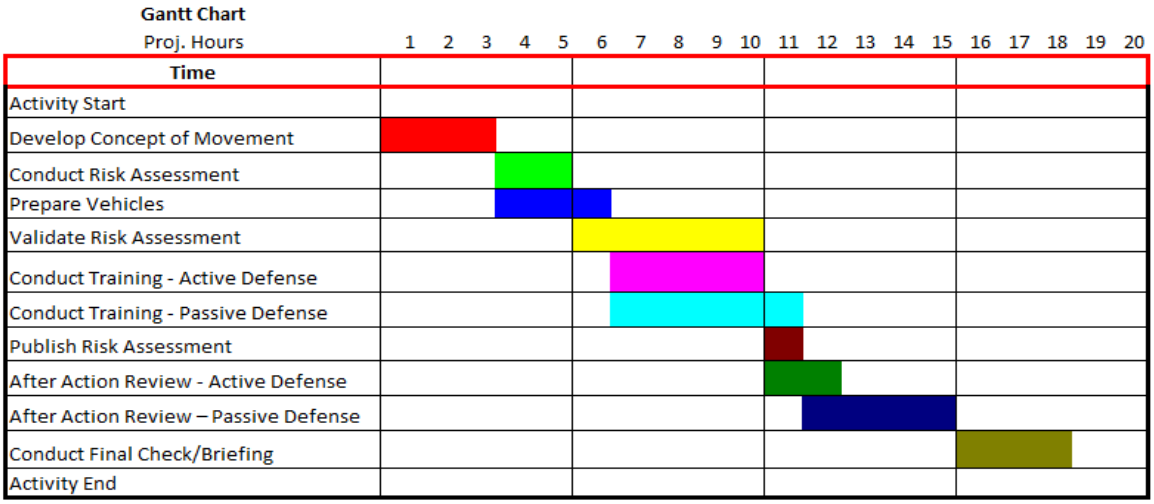
Fill in the table with your project data as shown below. Click the solve button that will be on the top left corner of the spreadsheet.

	Activity Data							Resources		
	Critical	Name	Description	Predecessors			Color	Res. 1	Time Hours	Activity Delay
1		Start	Activity Start					0	0	0
2		A	Develop Concept of Movement					0	3	0
3		B	Conduct Risk Assessment	A				0	2	0
4		C	Prepare Vehicles	A				0	3	0
5		D	Validate Risk Assessment	B				0	5	0
6		E	Conduct Training - Active Defense	C				0	4	0
7		F	Conduct Training - Passive Defense	C				0	5	0
8		G	Publish Risk Assessment	D				0	1	0
9		H	After Action Review - Active Defense	E				0	2	0
10		I	After Action Review – Passive Defense	F				0	4	0
11		J	Conduct Final Check/Briefing	G	H	I		0	3	0
12		End	Activity End					0	0	0
						11				

Once the network is solved, other buttons will appear. Click the Graph button to draw the network for your project.



A very useful tool for presentations and scheduling of the project is a Gantt chart. Click the Gantt button to create the chart as shown below.



4. PERT Example.

a. PERT uses uncertainty within time units. Known probability distributions for the activity times may be used. However, if not known, subject matter experts can be questioned to obtain optimistic, pessimistic and most likely completion times for each activity. A weighted scheme is then used to determine the activity mean and variance as follows:

$$Mean = \frac{a + 4m + b}{6}$$

$$Variance = \frac{(b - a)^2}{36}$$

b. Let's consider the following example of a computer terminal installation:

Activity	Task	Predecessor(s)	Time (Wks) Optimistic a	Time (Wks) Most Likely m	Time (Wks) Pessimistic b
Prepare Site	A	None	2.0	3.0	5.0
Order/Receive Terminals	B	None	2.8	3.7	6.2
Order/Receive Supplies	C	None	0.6	1.7	3.2
Install Telecommunications	D	A	1.8	2.8	4.8
Train Terminal Operators	E	A	2.9	3.1	5.3
Test/Correct Equipment	F	B, D	0.8	1.8	3.8
Seminars and Staff Training	G	B, D	2.2	4.2	6.0
Train Programmers	H	C, F	4.2	4.8	6.4
Write Program Packages	I	H	8.8	11.8	15.8

c. Using the above formulas, we calculate the mean and variance for each activity. For example, the calculations for activity A would be as follows:

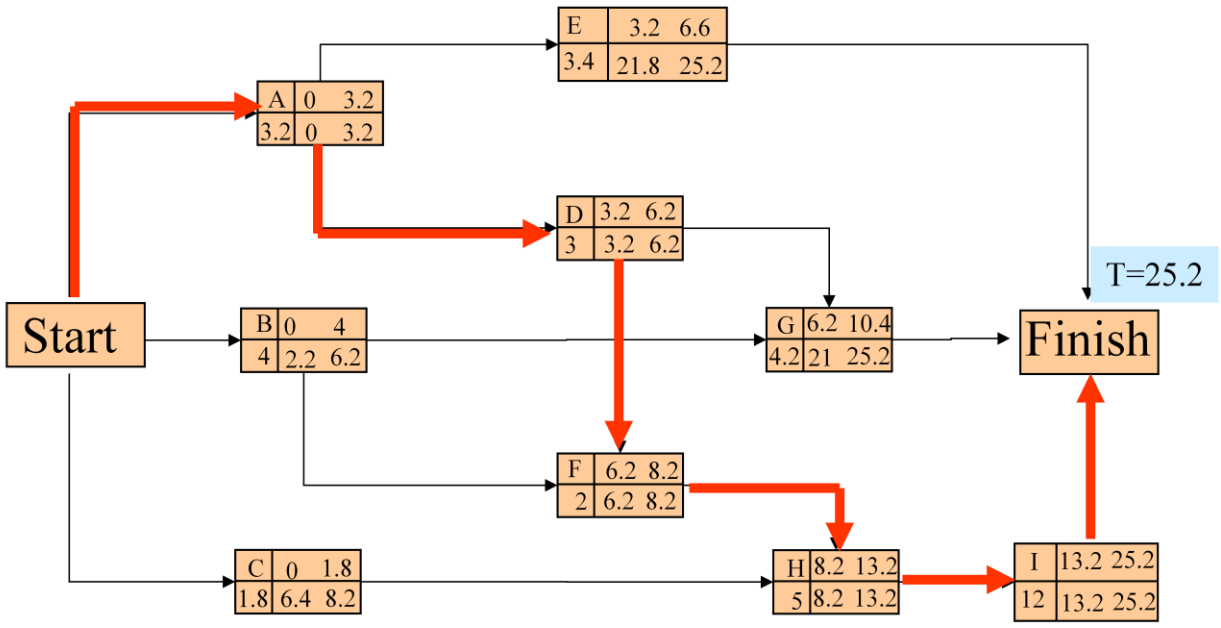
$$Mean_A = \frac{a + 4m + b}{6} = \frac{2 + 4(3) + 5}{6} = 3.167 \approx 3.2$$

$$Variance_A = \frac{(b - a)^2}{36} = \frac{(5 - 2)^2}{36} = 0.25$$

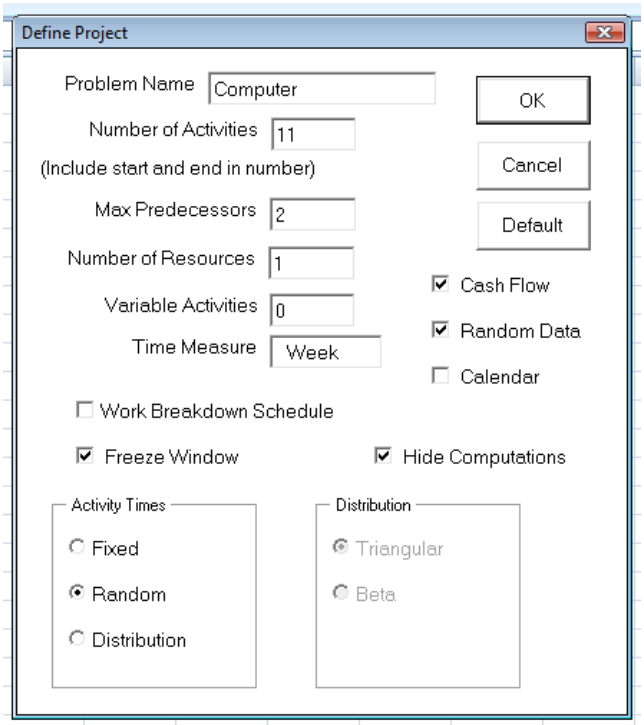
After completing each computation you would get the following results:

Task	Mean	Variance
A	3.2	0.25
B	4.0	0.32
C	1.8	0.19
D	3.0	0.25
E	3.4	0.16
F	2.0	0.25
G	4.2	0.40
H	5.0	0.13
I	12.0	1.36

d. We can now use the mean times of each project to create the network as we did with CPM. We would get the following network:



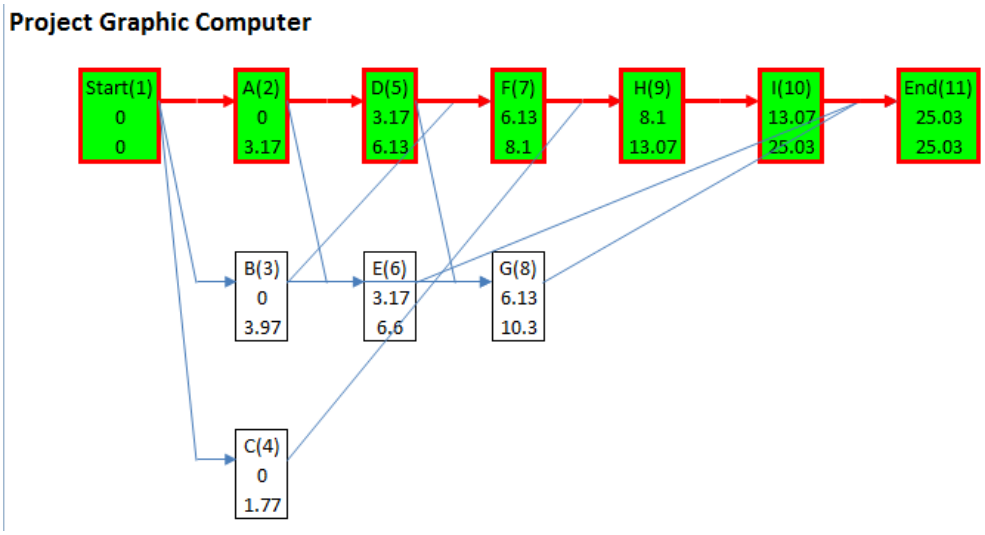
e. When using Jensen's add-in, make sure to highlight the Random function under Activity Times. This will give you the input columns for optimistic, pessimistic and most likely times.



The table inputs would look as follows:

	Activity Data						Resour	Activity Time Estimates				
	Critical	Name	Description	Predecessors		Color	Res. 1	Min.	Most Likely	Max.	Time Weeks	Activity Delay
1		Start	Activity Start				0	0	0	0	0	0
2		A	Prepare Site				0	2	3	5	3.2	0
3		B	Order & receive Terminals				0	2.8	3.7	6.2	4.0	0
4		C	Order & receive Supplies				0	0.6	1.7	3.2	1.8	0
5		D	Install telecommunications	A			0	1.8	2.8	4.8	3.0	0
6		E	Train terminal Operators	A			0	2.9	3.1	5.3	3.4	0
7		F	Test & correct equipment	B	D		0	0.8	1.8	3.8	2.0	0
8		G	Seminars & staff training	B	D		0	2.2	4.2	6	4.2	0
9		H	Train programmers	C	F		0	4.2	4.8	6.4	5.0	0
10		I	Write program packages	H			0	8.8	11.8	15.8	12.0	0
11		End	Activity End				0	0	0	0	0	0
					9							

And the network:



Note that the end project mean time is slightly different than the one we had calculated. That is due to using rounding in our calculations, where Excel holds all values within the calculations so it's more precise.

f. If applying the assumption that the critical path is normally distributed, then one can calculate the probability of completing a project with a desired time period using the cumulative distribution function for the normal random variable.

(1) First calculate the project standard deviation by adding the variances of the activities along the critical path and taking the square root.

$$\sigma_T = \sqrt{\text{sum of activitiy variances on critical path}}$$

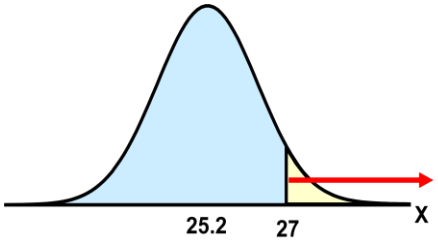
For this project the standard deviation would be:

$$= \sqrt{0.25 + 0.25 + 0.25 + 0.13 + 1.36} = 1.50$$

(2) With this information, you can answer probability questions as discussed in Section Four. For example, you boss may be interested in knowing the chances of over running the project by 27 weeks.

$$\text{Prob}(\text{time} \le \#\text{ of days}) = \text{Prob}\left(z \le \frac{(\#\text{ of days - T})}{\sigma_T}\right)$$

In this question we'd be looking for the area to the right of the curve of 27 weeks.



Given: $T = 25.2,\ \sigma_T = 1.50$

Then: $Z = \dfrac{(27 - T)}{\sigma_T} = \dfrac{(27 - 25.2)}{1.50} = 1.20$

Finding 1.20 in the normal distribution table yields a value of 0.8849. Therefore:

P(X ≤ 27) = P(Z ≤ 1.20) = 0.8849

P(X ≥ 27) = P(Z ≥ 1.20) =1-P(Z ≤ 1.20)= 1 – 0.8849 = 0.1151

Probability that the project will take more than 27 weeks is 11.51%.

If the boss is not willing to take that risk, then you would have to go into the project and see how to reduce times and variances by perhaps applying additional resources to the tasks.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884

6. Project management is a good tool to understand a project, its tasks, and potential risks prior to actually beginning the project.

7. For additional reading material on Project Management, consider

a. Practical Management Science, Rev. 3e, by Winston & Albright (ISBN: 0-324-66250-5). Chapter 15.2, the Basic CPM Model, provides illustrative examples and challenging exercises.

b. Operations Research – Applications and Algorithms, 4th Edition by Winston (ISBN: 0-534-38058-1) Chapter 8.4 blends CPM and PERT analysis together well.

c. Excel Add-In: www.me.utexas.edu/~jensen/ORMM/excel/project.html

SECTON NINE

MATH PROGRAMMING

(Return to Table of Contents)

1. Introduction.

   a. Math programming is a modeling approach that searches among many possible alternatives to find the "optimal" solution while attempting to allocate scarce resources among competing activities.

   b. There are several categories of math programming. These are:

      1) Linear Programming. This has a single linear objective function, subject to linear constraints.

      2) Integer Programming. This is a linear program that requires an integer solution. For this type of program a constraint is set that the solution must be integers (whole numbers).

      3) Goal Programming. This type has more than one linear goal, subject to linear constraints. Typically the goals are placed in priority order and solved as a linear program one at a time. As a solution is found for one goal, the next goal is solved with the solution set of the previous goal as a constraint.

      4) Non-linear Programming. In this type, objectives, goals and/or constraints can be non-linear.

   c. Since we will be discussing linear programming in this section, let's take a look at what are considered linear functions. A linear function is an algebraic expression with constants and variables. In math programming we place the constant (usually our constraint or requirement) to the right side of the equation and the variable/s to the left side. The equations can be equalities or inequalities. Linear expressions are not limited on the number of variables. Example functions may look like these:

$$3x + 2y = 10$$
$$3x_1 + 4x_2 + 2x_3 \leq 56$$
$$3y \geq 15$$

   Functions with an equal sign require solutions on the linear object (line, plane, etc.). Those with inequalities have solutions within an area above or below the linear object.

d. Nonlinear functions are expressions that do not form linear objects. A nonlinear function is an expression where the variables are multiplied together or variables have exponents other than 0 or 1. Examples include:

$$3xy + 2y = 250$$
$$3x + 4y^2 \leq 44$$
$$2x^{-1} + 3y \geq 35$$

2. Linear Math Programming. Linear programming has one objective function, subject to a series of constraints where all math statements must be linear and fractional answers to the decision variables are possible. To develop a linear program:

   a. Determine Objective. Determine the objective you must meet. Typically you will either minimize or maximize some criteria. For example: Minimize the cost of purchasing a system, Maximize the effects on a target.

   b. Determine the Decision Variables. Based on the objective, determine the variables that influence the decision. For example: To minimize cost of purchasing a system, the number of systems purchased influences the decision.

   c. Develop Objective Function. Using the objective and variables, develop the function to optimize. This will be a linear algebraic expression.

   d. Develop Constraints. Based on the objective, determine the scarce resources or minimum requirements that will hinder you reaching the objective. For example: a budget, a number of trucks available, or a performance requirement. The constraints will also be developed as linear algebraic expressions.

   e. To continue our study, let's use the following example: Your unit wishes to maximize the amount of cargo it can deliver on a daily basis. You can buy two types of vehicles for its delivery fleet. Each vehicle type has different operational costs. Each also has maintenance requirements in two shops. What mix of vehicles should it purchase?

      1) First we would have to collect the data for each vehicle. Let's say the data is as follows:

	Type 1	Type 2	Resources Avalaible
Operational Cost	$2K	$5K	$60K
Manhours, Shop A	2hr	1hr	40hr
Manhours, Shop B	5hr	2hr	40hr
Total Carrying Capacity	5K tons	10K tons	

2) The next step is to determine what the variables are. The objective is to maximize the amount of cargo the unit can carry. The way to accomplish the task is by purchasing two types of vehicles. Therefore the information we need to seek is: the amount of cargo that can be carried and the amount of each type vehicle to purchase. The variable Z, is conventionally used in math programming as the objective. Therefore the variables we have are:

T_1 = The number of Type 1 trucks bought
T_2 = The number of Type 2 trucks bought
Z = The total carrying capacity of the Vehicle fleet

3) Next, we develop the objective function using the variables. Since we want to maximize the cargo carried the function has to use the capacity of each vehicle. Therefore our function is:

$$\text{Max } Z = 5\ T_1 + 10\ T_2$$

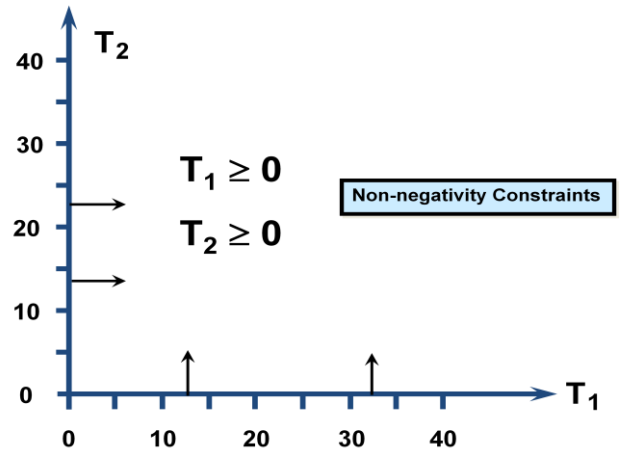
4) We now develop our constraints. We have three constraints, the two maintenance requirements, and the operational costs. We need a function for each one. In each case, we have a resource we cannot exceed, therefore our function is an inequality of the form less than or equal to, $\leq$. The three functions look like this:

$2\ T_1 + 5\ T_2 \leq 60$ Operational Cost
$2\ T_1 + 1\ T_2 \leq 40$ Shop A
$5\ T_1 + 2\ T_2 \leq 40$ Shop B

There are two more implied constraints in many math programming situations. These are called non-negativity constraints. We do not desire nor able to purchase a negative number of vehicles. Therefore:

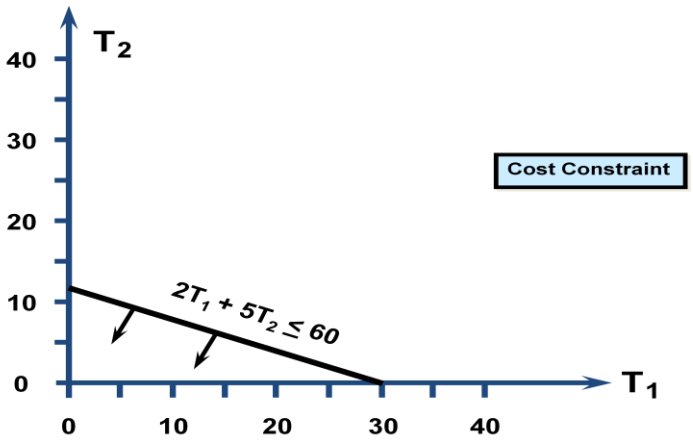
$T_1 \geq 0$ and $T_2 \geq 0$ Non-negativity constraint

5) We can now begin to solve our problem. There are several techniques to solve math programming models. The most common is a system of linear equations technique called the simplex method. For simple problems consisting of two variables, the problem can be solved graphically. Of course, most problems are solved using computer models. We will solve our problem graphically first to demonstrate the idea of what is involved. The way to do this is by graphing the constraints to find an area where a solution lies. The first constraints to graph are the non-negativity constraints, which limits the solution set to the positive quadrant of the graph. This is shown on the graph at the top of the next page.



Graph of Non-Negativity Constraints

Next we graph each constraint.



Graph of Cost Constraint

Because we are dealing with inequalities, the graph of each constraint is represented by an area bound by a line. As a refresher on graphing, let's demonstrate how we graphed the cost constraint $2T_1 + 5T_2 \leq 60$. The best way is to find each intercept point (i.e. the locations of the x-axis intercept and the y-axis intercept). This is done by setting each variable to zero, one at a time, and solving for the other variable. Setting $T_2 = 0$ we get:

$2T_1 + 5(0) = 60$
$2T_1 = 60$

Dividing each side by 2 to solve for T_1 results in $T_1 = 30$

Now plot the point (30, 0) as shown in the graph above.

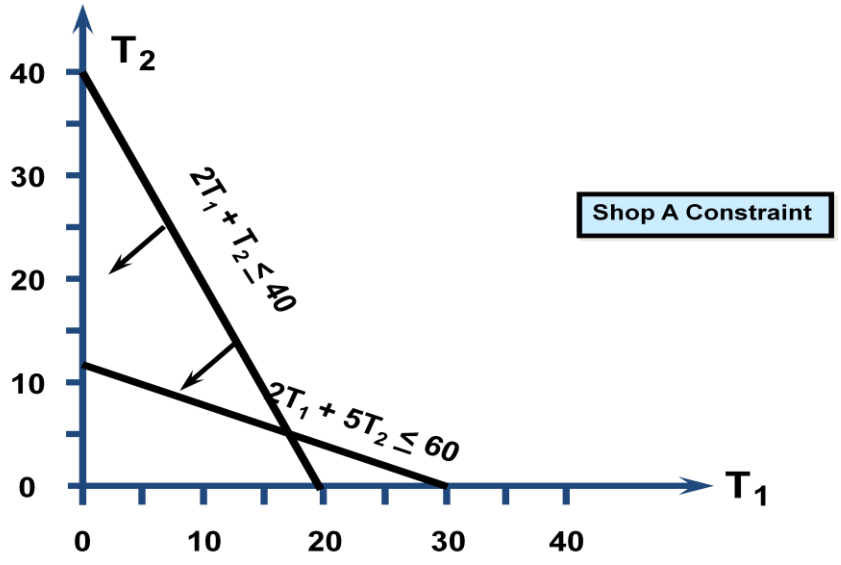
Do the same for the variable T_2 as shown below.

$2(0) + 5T_2 = 60$
$5T_2 = 60$

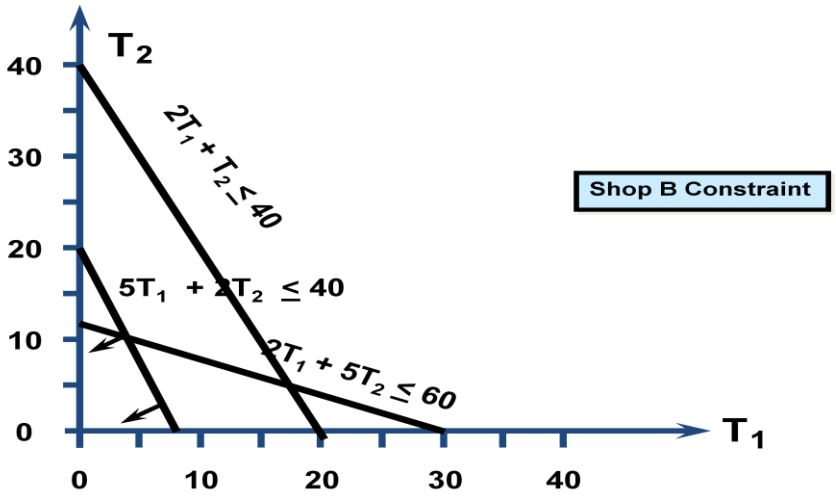
Dividing each side by 5 to solve for T_2 results in $T_2 = 12$

Now plot the point (0, 12) and draw a line between (30, 0) and (0, 12) to get the boundary line represented by the equation $2T_1 + 5\ T_2 = 60$. Since the inequality is a less than constraint, the area corresponding to the cost constraint is the area under the line as indicated by the arrows. If you are unsure as to which way the arrows should point, pick an arbitrary point on each side of the boundary line to see which one satisfies the inequality. The arrows point in the direction of that point. For example, the point (0, 0) located below the boundary line satisfies the inequality (i.e. 2(0) + 5(0) = 0 which is less than 60. On the other hand, the point (20, 20) located above the boundary does not satisfy the inequality (i.e. 2(20) + 5 (20) = 140 which is greater than 60. Therefore, the arrows corresponding to the cost constraint should point down towards the origin as depicted in the graph.

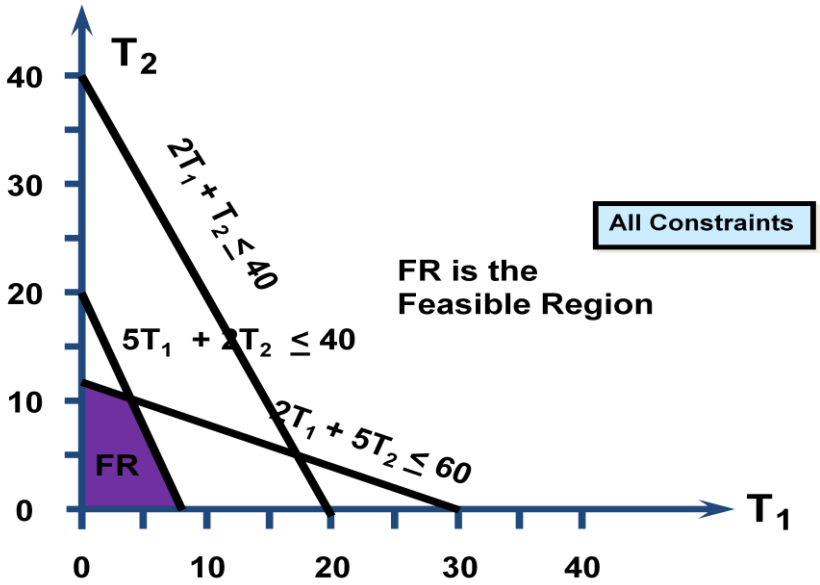
Add each constraint to the graph in turn.



Graph with Shop A Constraint Added

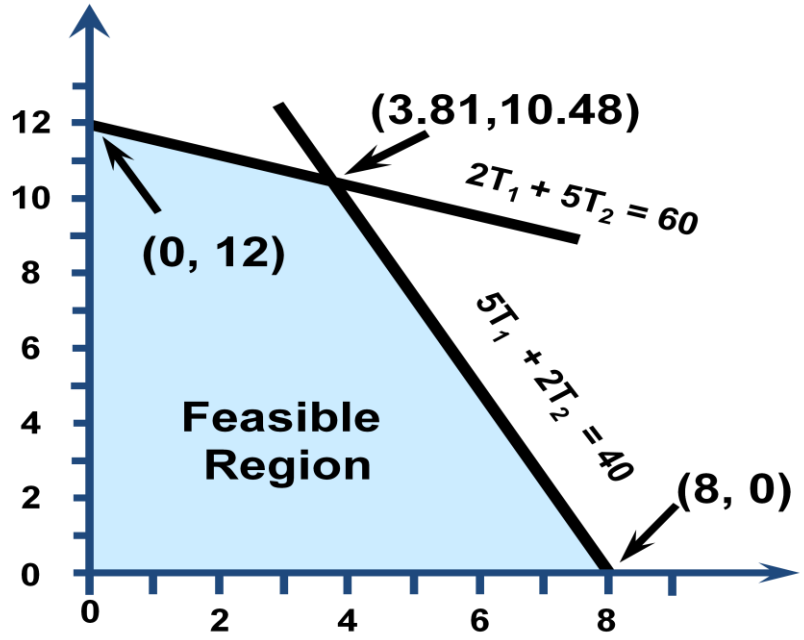


Graph with Shop B Constraint Added



Graph of the Feasible Region

The shaded area is called the feasible region and represents all possible solutions that will satisfy all of the constraints simultaneously. A blowup of this region is shown in the graph at the top of the next page. We now have to find a solution in this area that will provide us with the greatest amount of carrying capacity. The boundaries of the feasible region are defined by the corner points (0, 0), (0, 12), (8, 0), and (3.8, 10.47). The first three points correspond to the axes intercept points and quadrant boundaries. The last point (3.8, 10.47) is the intersection of the two lines $2T_1 + 5T_2 = 60$ and $5T_1 + 2T_2 = 40$. We can find this point by solving the two equations simultaneously. Again, as an algebra review, let's see how this is done.



Blowup of the Feasible Region

One algebra method commonly used is the substitution method. Select one of the equations and solve for one of the variables in terms of the other. Let's take the equation $2T_1 + 5T_2 = 60$ and solve for T_1.

$$2T_1 + 5T_2 = 60$$
$$2T_1 = 60 - 5T_2$$
$$\frac{2T_1}{2} = \frac{60}{2} - \frac{5T_2}{2}$$
$$T_1 = 30 - \frac{5}{2}T_2$$

Now substitute the expression for T_1 in the equation $5T_1 + 2T_2 = 40$ and solve for T_2.

$$5T_1 + 2T_2 = 40$$
$$5\left(30 - \frac{5}{2}T_2\right) + 2T_2 = 40$$

Multiplying through by 5, results in the following equation.

$$150 - \frac{25}{2}T_2 + 2T_2 = 40$$

Combining like terms we get

$$150 - \frac{21}{2}T_2 = 40$$

Now solve for T_2. First subtract 150 from both sides to get

$$-\frac{21}{2}T_2 = -110$$

Multiply both sides of the equation by $-\frac{2}{21}$ to get

$$T_2 = \frac{220}{21} = 10.47.$$

Substitute this value for T_2 in the equation $T_1 = 30 - \frac{5}{2}T_2$ and solve for T_1.

$$T_1 = 30 - \frac{5}{2}T_2 = 30 - \frac{5}{2}(10.47) = 3.8.$$

In linear programming the optimal solution will always be at one of the corner points. To find the best solution that maximizes cargo, it makes sense that we want to purchase the most amount of vehicles possible. Therefore, we want to be as far away for (0, 0) as possible. We can do this 3 ways; along the T_1 axis, along the T_2 axis, or along both at the same time. If we do that, the furthest points will be the three corner points. That is what a math program will do. It will seek a solution along the corner points, eliminating the infinite possibilities within the area. So if we take the objective function, Max Z = 5 T_1 + 10 T_2, we substitute each corner point and find the best solution.

Max Z = 5 T_1 + 10 T_2
At corner point (8, 0): 5 (8) + 10 (0) = 40
At corner point (0, 12): 5 (0) + 10 (12) = 120
At corner point (3.8, 10.47): 5 (3.8) + 10 (10.47) = 123.8

The best solution is to purchase 3.8 Type 1 vehicles and 10.47 Type 2 vehicles with a carrying capacity of 123.8 tons. Remember that math programming allows for partial solutions. Of course we cannot purchase decimals of vehicles so this solution is not valid for this type of problem. However, if we were talking about time we would not have a problem with 3.8 days. The next step is to then find a solution that will give us whole vehicles. This is called integer programming.

3. Integer program. Integer programming is a type of linear program that requires the solution set to be integers (whole numbers). The technique is to conduct a search for integer sets near the optimal solution within the feasible region.

    a. You cannot always simply round the optimal solution because that could lead you to be outside the feasible region.

b. Looking back to the corner points, we found one integer solution, (0, 12), with a capacity of 120 tons. Now, let's see if there may be another solution set that does any better. If we move down the T_2 line to 11, we find we can move as far as 2 on the T_1 line. If we were to have moved to (3, 11), we would have not been in the feasible region. Substituting the values 3 and 11 into cost constraint results in a cost greater than 60.

$$2\ T_1 + 5\ T_2 = 2(3) + 5(11) = 61$$

Now we can evaluate (2, 11) in the objective function.

$$\text{Max } Z = 5\ T_1 + 10\ T_2\ \ 5\ (2) + 10\ (11) = 120.$$

Note that this set of values yields the same solution as (0, 12). Let's set $T_2 = 10$. At this value we can move out to $T_1 = 4$. This also results in a maximum carrying capacity of 120 (i.e. 5(4) + 10(10) = 120). Note that the point (5, 10) is outside the feasible region.

If we set $T_2 = 9$, the best we can do for T_1 without violating the feasible region is to set $T_1 = 4$, which obviously would result in a maximum carrying capacity less than the solution of $T_1 = 4$ and $T_2 = 10$.

So, in this problem we found three solution sets (0, 12), (2, 11), and (4, 10), all with a carrying capacity of 120 tons. Each one of these solutions satisfies our requirements and any can be selected. You may argue that the solution (0, 12) gives you one type of vehicle, thereby reducing training and adding commonality. Or, you may say that the solution (4, 10) offers 14 vehicles, allowing for more missions and more flexibility. Or even that one solution uses less of the budget. Although all these arguments are good, what you are doing is in fact adding more constraints to the problem. If you desire to consider all these factors, you need to add them to the problem upfront. When you solve this problem in a computer model, you will only see one solution. This may bother some of you, but also note that most problems are much more complex and therefore very difficult to find more than one "optimal" solution.

4. Computer models. There are many computer models for math programming. Most of the older models required the objective function and constraints to be algebraic inputs. Though, the algebraic functions are still present, Excel® takes a slighter different approach by organizing it in a spreadsheet. The add-in solver is used for math programming. A few key things to note when building a model in Excel: First, select one cell for the objective function. This cell requires a math formula in it. Second, you will have a set of blank cells for the solution input. These are your decision variables and will be input by the program. Also the formulas include these cells. Last, set up a group of cells for the constraints with appropriate formulas. We have developed a math programming template for use (see appendix B for download location).

a. Let's first look at the blank sheet. The sheet as pictured below only shows a few cells in each category. This model allows for 25 variables and 8 constraints in each category type, but can easily be expanded. The light green areas (row 3 and column A) are title cells. You can type in your variable names in row 3, such as T_1 or 5 Ton truck. In column A you can type in what the objective is and the constraints (i.e. Budget, Max tons). The yellow area (row 4) is for your decision variables, where the answer set will be filled in by the program. The objective function and constraint formulas are built in the blue areas in column AB. The cell AB7 is for the objective function. The formula in this cell will multiply the cell values found in row 7 (i.e. the objective function coefficients) by the decision variable values found in row 4. The other cells in column AB are for constructing the constraints and will multiply the row 4 decision variable values by the constraint coefficients found in rows 11 through 40.

	A	B	C	D	AB	AC	AD
1	**Linear Programming Model Template**						
2							
3							
4	**Dec var:**						
5							
6	**Objective:**				**OBJ value:**		
7					0		
8							
9							
10	**Less than or equal to constraints:**				**Constraint values:**		
11					0	<=	
12					0	<=	
13					0	<=	
19							
20	**Greater than or equal to constraints:**						
21					0	>=	
22					0	>=	
23					0	>=	
30							
31	**Equality constraints**						
32					0	=	
33					0	=	
34					0	=	

There are three types of constraints found in the model (≤, ≥, and =). There is a section in the model for each of these constraint types. Column AD is used to record the right-hand side values for the constraints. The model will compare the computed values in column AB with these right-hand side values to insure that the constraints are satisfied.

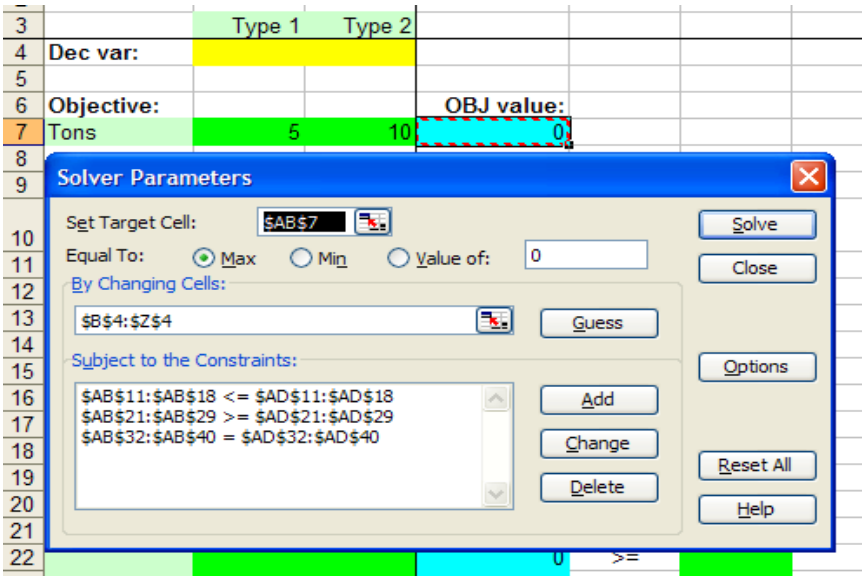
b. Now let's use the template to set up our example. First let's recall what our model looked like in algebraic terms.

$$\text{Max } Z = 5\,T_1 + 10\,T_2$$
s.t.
$2\,T_1 + 5\,T_2 \leq 60$ Operational Cost
$2\,T_1 + 1\,T_2 \leq 40$ Shop A
$5\,T_1 + 2\,T_2 \leq 40$ Shop B

We will use the coefficients of the algebraic model as the input into the spreadsheet. Because all the constraints are less than constraints, we only have to use the first part of the constraints in the spreadsheet. Note that row 4 is blank and the column AB values are all zero. Once solver finds the solution, it will fill out row 4 and the math will be completed in column AB. To solve this problem, first ensure that solver is activated as an add-in (see Appendix E).

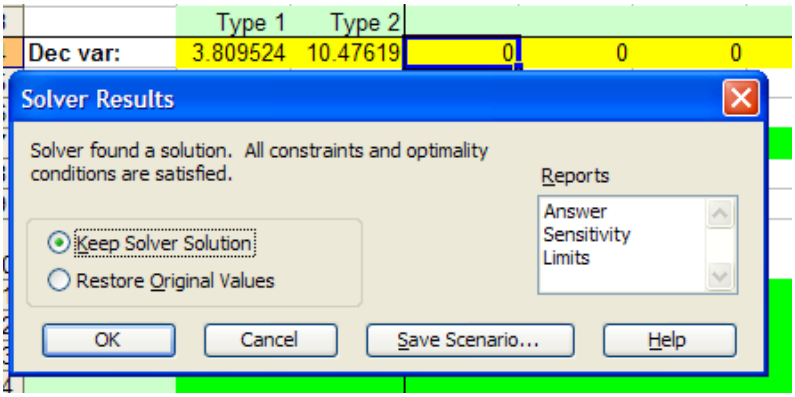
	A	B	C	AB	AC	AD
1	**Linear Programming Model Temp**					
2						
3		Type 1	Type 2			
4	**Dec var:**					
5						
6	**Objective:**			**OBJ value:**		
7	Tons	5	10	0		
8						
9						
10	**Less than or equal to constraints:**			**Constraint values:**		
11	Cost	2	5	0	<=	60
12	Shop A	2	1	0	<=	40
13	Shop B	5	2	0	<=	40

To call up Solver click on the tab for the Data ribbon and then click on Solver (located in the upper right-hand corner of the ribbon). You will get the following drop down menu.



Note that the cells of the menu are filled out. This is because the template is already setup. The first box, ***Set Target Cell***, is the Objective function value. Next you have 3 button options: ***Max, Min, or Value of***. We will only use ***Max or Min***. Make sure you select the right button for your problem. The next box, ***By Changing Cells***, is your decision variables. That is where you want your solution set. The last box, ***Subject to the Constraints***, is where you set up the constraints. Note there are three constraints. Each refers to one of the three types of constraints. We will show how to add a constraint later.

Clicking on the ***Solve*** button will return a solution and the box shown below. It is important to read the contents of the box. The statement, ***"Solver found a solution. All constraints and optimality conditions are satisfied"*** indicates that the program was able to derive an optimal feasible solution. You will also be given the opportunity to select several reports to print. At this point simply click on OK to retain the current solution.

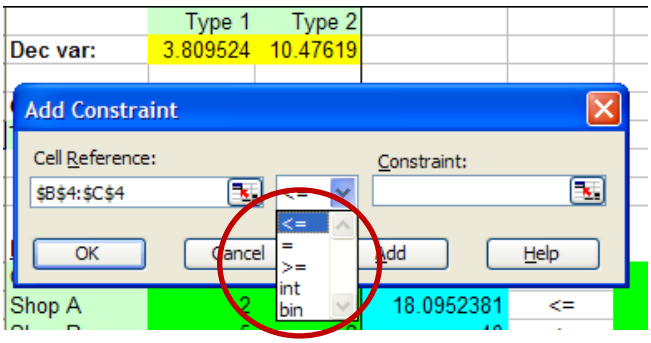


The program has filled in row 4 with the solution set and column AB with the objective and constraint values. As we found before the most capacity we can gain is 123.81 tons by purchasing 3.81 Type 1 and 10.48 Type 2 vehicles. We will use $60,000 in operational costs, 18.1 hours of Shop A maintenance and 40 hours of Shop B maintenance.

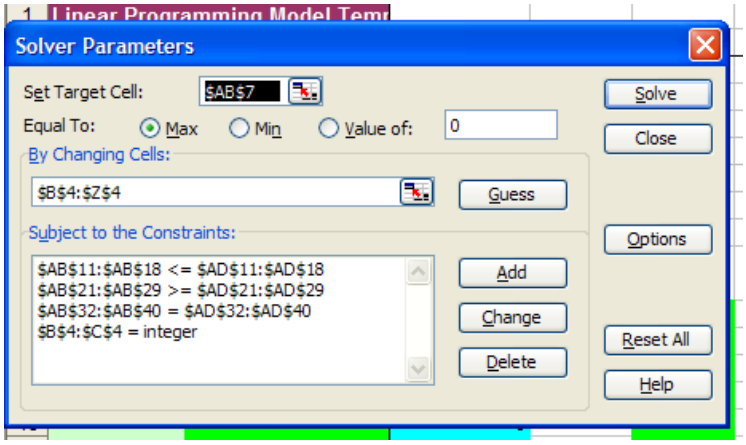
3		Type 1	Type 2			
4	**Dec var:**	3.809524	10.47619			
5						
6	**Objective:**			**OBJ value:**		
7	Tons	5	10	123.8095238		
8						
9						
10	**Less than or equal to constraints:**			**Constraint values:**		
11	Cost	2	5	60	<=	60
12	Shop A	2	1	18.0952381	<=	40
13	Shop B	5	2	40	<=	40

c. Sensitivity Analysis. At this point we can perform sensitivity analysis. We can look at changes that will affect the solution. Any change to the coefficients will change the solution. If you have some coefficients that you are not absolutely sure of, you may want to run the program with variations to see the affects. Remember to only change one variable at a time. For example, perhaps the operational cost of the Type 2 vehicle might vary from 4 to 7 thousand dollars. You can make these varying changes to see the affects. You can also change right-hand side values. Changes in binding constraints will always change the solution. For example, if you believe the budget may be cut to $55,000, then the solution will be affected. You can also look at it another way. If the decision maker is not satisfied with 123 tons, then you know you need an increase in either cost or Shop B hours. On the other hand an increase in Shop A has no affect since you are not using the allotted hours (i.e. using only 18 out of 40 available hours). You also have an allowable decrease to 19 hours in Shop B before the solution is affected.

d. As stated before, we cannot purchase partial vehicles. An integer solution is required. To accomplish this add a constraint to the model requiring the decision variables (Type 1 and Type 2) to be integer in value. To accomplish this click on the ***Solver*** button and ***Subject to Constraints*** click on the ***Add*** button. This action reveals the menu box shown below.



Place the cursor in the ***Cell Reference*** area and highlight cells B4 and C4. Then click on the down arrow (∨) to reveal the circled box. Click on ***int***. This tells Excel to treat the values in cells B4 and C4 as integers. Then click on OK. This will add the constraint $B$4:$C$4 = integer to the bottom of the list of constraints as in the figure below.

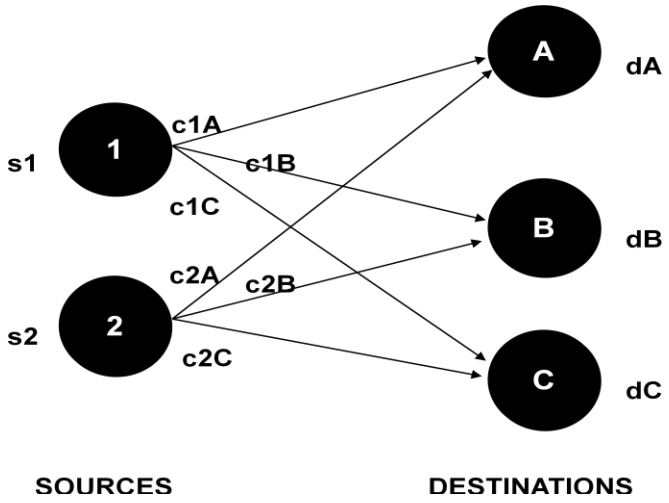


Now click on ***Solve***. Again, make sure Solver returns the statement that a solution was found. Now we have an integer solution that corresponds to one of the solutions found earlier. Note that you may have found one of the other possible integer solutions, depending on where your particular model began its search. Remember that most problems are more complex without multiple solutions.

3		Type 1	Type 2			
4	**Dec var:**	2	11			
5						
6	**Objective:**			**OBJ value:**		
7	Tons	5	10	120		
8						
9						
10	**Less than or equal to constraints:**			**Constraint values:**		
11	Cost	2	5	59	<=	60
12	Shop A	2	1	15	<=	40
13	Shop B	5	2	32	<=	40

5. Transportation Problems. We will discuss three basic types of transportation problems; Transportation, assignment, and transshipment problems. All these are types of networks. A network model is one which can be represented by a set of nodes, a set of arcs, and functions (e.g. costs, supplies, demands, etc.) associated with the arcs and/or nodes. They can all be formulated and solved by linear programs. There are also many computer packages that are specific for more complex transportation problems.

a. Transportation Problems. These problems require the movement of some product from a number of sources with limited supply to a number of destinations with specified demand. The usual criterion is to minimize total transportation cost (monetary, mileage, or time). Shown at the top of the next page is a network representation of a transportation problem. In this general network, there are two sources of supply, 1 and 2, with supply quantities of s1 and s2. There are three destinations, 1, 2 and 3, that require amounts d1, d2, and d3 respectively. The cost of sending one unit of supply to the respective destination is represented by c11, c12, etc. The objective of a transportation problem is to minimize cost.



So it follows that the objective function would be to minimize the sum of the costs. There are two types of constraints to deal with; the amount of supply each source has available and the amount of supply each destination requires. The general math formulation of a transportation problem follows:

Let x_{ij} = the amount shipped from source i to destination j.

$$\text{Min} \sum_i \sum_j c_{ij} x_{ij} \qquad \text{(minimize total cost, time, or distance)}$$

s.t.

$$\sum_j x_{ij} \le s_i \quad \text{for each source i (set of supply constraints)}$$

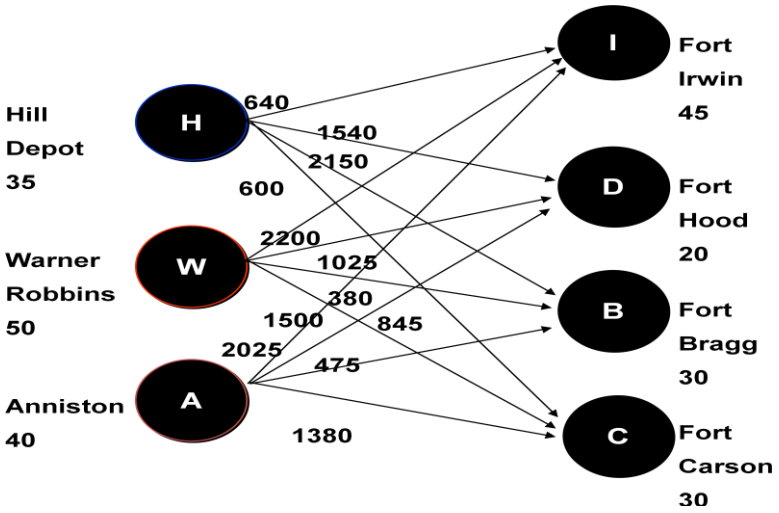
$$\sum_i x_{ij} \ge d_j \quad \text{for each destination j (set of demand constraints)}$$

$$x_{ij} \ge 0 \quad \text{for all i and j (set of non-negativity constraints)}$$

b. Let's look at an example: Fort Bragg, Hood, Irwin, and Carson require weekly delivery of a certain commodity, as described in the following table. Hill Depot, Warner Robbins, and Anniston can all deliver the commodity to all locations; however are limited in the amount of supply they each have. Construct a linear program to deliver the supplies with the shortest distance possible.

	Destinations				
Source	**Fort Irwin**	**Fort Hood**	**Fort Bragg**	**Fort Carson**	**Supply (K tons)**
Hill Depot	**640 miles**	**1540 miles**	**2150 miles**	**600 miles**	**35**
Warner Robbins	**2200 miles**	**1025 miles**	**380 miles**	**1500 miles**	**50**
Anniston	**2025 miles**	**845 miles**	**475 miles**	**1380 miles**	**40**
Demand (K tons)	**45**	**20**	**30**	**30**	

The First thing you may want to do is add the amount of available supply and the amount of required demand to make sure there is sufficient supply to meet demand. We will discuss what to do if supply does not meet demand. In this case the total amount of available supply, 125 tons, equals the total demand of 125 tons. Second, you want to create the network. This will assist you with the formulation of the math program. Essentially, each arrow within the network represents a decision variable and each node a constraint.



The next step is to formulate the problem algebraically. We'll use the following notation to represent the different locations.

H = Hill Depot
W = Warner Robbins
A = Anniston
I = Fort Irwin
D = Fort Hood
B = Fort Bragg
C = Fort Carson

$$\text{Min}: Z = 640x_{HI} + 1540x_{HD} + 2150x_{HB} + 600x_{HC} + 2200x_{WI} + 1025x_{WD} + 380x_{WB} + 1500x_{WC} + 2025x_{AI} + 845x_{AD} + 475x_{AB} + 1380x_{AC}$$

s.t.

Supply

$x_{HI} + x_{HD} + x_{HB} + x_{HC} \le 35$

$x_{WI} + x_{WD} + x_{WB} + x_{WC} \le 50$

$x_{AI} + x_{AD} + x_{AB} + x_{AC} \le 40$

Demand

$x_{HI} + x_{WI} + x_{AI} \ge 45$

$x_{HD} + x_{WD} + x_{AD} \ge 20$

$x_{HB} + x_{WB} + x_{AB} \ge 30$

$x_{HC} + x_{WC} + x_{AC} \ge 30$

$x_{ij} \ge 0$ for all i and j.

We now place this formulation into the Excel math programming template and solve.

Linear Programming Model Templa															
	Hill to Irwin	Hill to Hood	Hill to Bragg	Hill to Carson	Warner to Irwin	Warner to Hood	Warner to Bragg	Warner to Carson	Anniston to Irwin	Aniston to Hood	Anniston to Bragg	Anniston to Carson			
Dec var:	35	0	0	0	0	0	30	20	10	20	0	10			
Objective:													OBJ value:		
Miles	640	1540	2150	600	2200	1025	380	1500	2025	845	475	1380	114750		
Less than or equal to constraints:													Constraint values:		
Hill Supply	1	1	1	1									35	<=	35
Warner Supply					1	1	1	1					50	<=	50
Anniston Supply									1	1	1	1	40	<=	40
Greater than or equal to constraints:															
Irwin Demand	1				1				1				45	>=	45
Hood Demand		1				1				1			20	>=	20
Bragg Demand			1				1				1		30	>=	30
Carson Demand				1				1				1	30	>=	30

Therefore, the solution is for Fort Irwin to receive 35 tons from Hill Depot, 10 tons from Anniston, Fort Hood to receive 20 tons from Anniston, Fort Bragg to receive 30 tons from Warner Robbins and Fort Carson to receive 20 tons from Warner Robbins, 10 tons from Anniston for a total of 114,750 ton miles.

c. Now let's look at some adjustments that can be made to a transportation problem.

1) First, if supply does not meet demand, that is, there is a shortage of supply. For the model to work, the formulation must be balanced. To do this, create a 'dummy' shipper. The 'dummy' will have no cost to ship (0 as the decision coefficient), and have a supply quantity equal to the shortage amount. The amount 'shipped' from the 'dummy' will be the amount owed the destination.

2) You may have unacceptable routes. Causes for unacceptable routes can be blocked roads, air leg beyond aircraft reach, political differences between source and destination, etc. Unacceptable routes are shown in the network by the elimination of the route (arrow), and in the model by eliminating the variable.

3) You may also have to consider route capacity. Some routes may only be able to handle a specific amount of supply. This is accomplished by adding a constraint to that route.

4) Finally, you may consider minimum shipping guarantees. In the case of supply shortages, you may want to ensure all destinations receive at least some minimum amount. This is done by adding a constraint to the model requiring the minimum shipping.

d. Assignment Problems. An assignment problem seeks to minimize the total cost assignment of *m* agents to *n* tasks, given that the cost of agent *i* performing task *j* is c_{ij}. It is a special case of a transportation problem in which all supplies and all demands are equal to 1; hence assignment problems may be solved as linear programs.

The general formulation of an assignment problem is a follows:

Let x_{ij} = 1 if agent i is assigned to task j, 0 otherwise.

$$\text{Min} \sum_i \sum_j c_{ij} x_{ij} \qquad \text{(minimize total cost or time)}$$

s.t.

$$\sum_j x_{ij} \le 1 \quad \text{for each agent i (each agent performs only 1 task)}$$

$$\sum_i x_{ij} \ge 1 \quad \text{for each task j (each task requires one agent)}$$

$$x_{ij} \ge 0 \quad \text{for all i and j (set of non-negativity constraints)}$$

e. Let's work the following example: During the next mission, the division requires four tasks to be completed requiring the use of aircraft. The unit has six aircraft available. Each aircraft can perform each mission, however each require different flight times for each mission as shown in the following table. To minimize risk the G3 is interested in assigning the aircraft to minimize flight time.

	Missions			
Aircraft	**1**	**2**	**3**	**4**
A	**18**	**13**	**17**	**14**
B	**16**	**15**	**16**	**15**
C	**14**	**14**	**20**	**17**
D	**20**	**13**	**15**	**18**
E	**16**	**18**	**15**	**17**
F	**20**	**15**	**13**	**16**

The problem formulation would look as follows:

Let x_{ij} = 1 if agent i is assigned to task j, 0 otherwise

Min Z = $18x_{A1} + 13x_{A2} + 17x_{A3} + 14x_{A4} + 16x_{B1} + 15x_{B2} + 16x_{B3} + 15x_{B4} + 14x_{C1} + 14x_{C2} + 20x_{C3} + 17x_{C4} + 20x_{D1} + 13x_{D2} + 15x_{D3} + 18x_{D4} + 16x_{E1} + 18x_{E2} + 15x_{E3} + 17x_{E4} + 20x_{F1} + 15x_{F2} + 13x_{F3} + 16x_{F4}$

s.t.

Aircraft Assigned

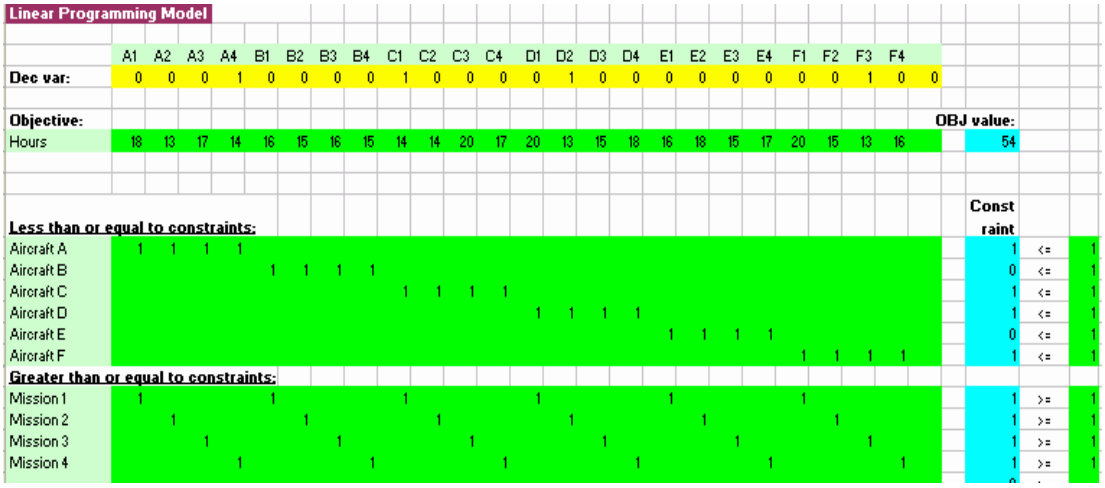
$x_{A1} + x_{A2} + x_{A3} + x_{A4} \leq 1$
$x_{B1} + x_{B2} + x_{B3} + x_{B4} \leq 1$
$x_{C1} + x_{C2} + x_{C3} + x_{C4} \leq 1$
$x_{D1} + x_{D2} + x_{D3} + x_{D4} \leq 1$
$x_{E1} + x_{E2} + x_{E3} + x_{E4} \leq 1$
$x_{F1} + x_{F2} + x_{F3} + x_{F4} \leq 1$

Missions Assigned

$x_{A1} + x_{B1} + x_{C1} + x_{D1} + x_{E1} + x_{F1} \geq 1$
$x_{A2} + x_{B2} + x_{C2} + x_{D2} + x_{E2} + x_{F2} \geq 1$
$x_{A3} + x_{B3} + x_{C3} + x_{D3} + x_{E3} + x_{F3} \geq 1$
$x_{A4} + x_{B4} + x_{C4} + x_{D4} + x_{E4} + x_{F4} \geq 1$

$x_{ij} = 0$ or 1 for all i and j

The template would look as below with the following result: Aircraft A assigned mission 4, Aircraft C to mission 1, Aircraft D to mission 2, and Aircraft F to mission 3. Aircrafts B and E would not be required. A total of 54 flight hours would be spent.



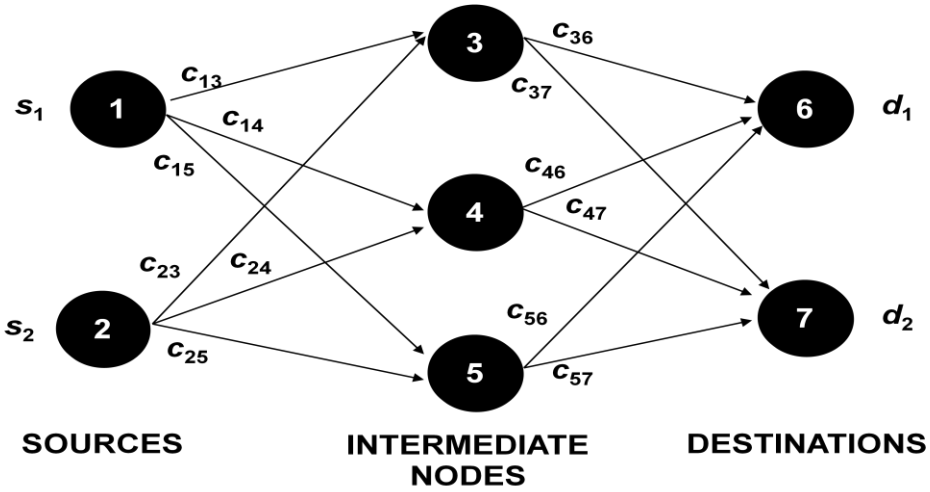
Linear Programming Model																														
	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4	D1	D2	D3	D4	E1	E2	E3	E4	F1	F2	F3	F4						
Dec var:	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0					
Objective:																											OBJ value:			
Hours	18	13	17	14	16	15	16	15	14	14	20	17	20	13	15	18	16	18	15	17	20	15	13	16			54			
Less than or equal to constraints:																											Constraint			
Aircraft A	1	1	1	1																							1	<=	1	
Aircraft B					1	1	1	1																			0	<=	1	
Aircraft C									1	1	1	1															1	<=	1	
Aircraft D													1	1	1	1											1	<=	1	
Aircraft E																	1	1	1	1							0	<=	1	
Aircraft F																					1	1	1	1			1	<=	1	
Greater than or equal to constraints:																														
Mission 1	1				1				1				1				1				1						1	>=	1	
Mission 2		1				1				1				1				1				1					1	>=	1	
Mission 3			1				1				1				1				1				1				1	>=	1	
Mission 4				1				1				1				1				1				1			1	>=	1	

f. As with transportation problems, some modifications can be discussed with assignment problems.

1) Unacceptable agent to task. An agent may not be able to perform a specific task, i.e., a task requires cargo aircraft and one of the agents is an attack aircraft.

2) Agents doing more than one task. You may allow agents to perform more than one task. In this case, it is important how you define the decision variable and how you want the agents to perform. For example, you may allow aircraft A to perform 2 missions. If you define decision variables as time to complete mission and return and you want the aircraft to return prior to the next mission, then simply add a 2 for aircraft A. If you desire aircraft to complete a mission and continue to the next mission, then you must create variables with associated costs between all the possibilities.

g. Transshipment Problems. Transshipment problems are transportation problems in which a shipment may move through intermediate nodes (transshipment nodes) before reaching a particular destination node. A general network of transshipment problem is displayed at the top of the next page. As represented the supply from the source nodes must travel through one of the intermediate nodes before reaching the final destination. There are several reasons for transshipment nodes – refueling, changing to a different mode of transportation, breaking or consolidating cargo, immigration, etc.



The general formulation of the transshipment problem follows.

Let x_{ik} = the amount shipped from node i (source) to node k (intermediate)
Let x_{kj} = the amount shipped from node k (intermediate) to node j (destination)

$$\text{Min} \sum_i \sum_k c_{ik} x_{ik} + \sum_k \sum_j c_{kj} x_{kj} \qquad \text{(minimize total cost, time, or distance)}$$

s.t.

$$\sum_k x_{ik} \le s_i \qquad \text{for each origin or source i}$$

$$\sum_k x_{ik} - \sum_j x_{kj} = 0 \qquad \text{for each intermediate node k}$$

$$\sum_j x_{kj} \ge d_j \qquad \text{for each destination j}$$

$$x_{ik} \ge 0 \text{ and } x_{kj} \ge 0 \qquad \text{for all i, j, and k}$$

Note that the objective function includes the costs of going from the source to the intermediate nodes and the cost of going from the intermediate nodes to the destination nodes. The supply constraints take the supply to the intermediate nodes. The demand constraints receive the shipments from the intermediate nodes. At the intermediate nodes there must be conservation of flow. That is, the flow in must be exactly equal to the flow out and thus, the right-hand side values for these constraints

must be zero.

h. Let's look at an example. The Army needs to ship a certain commodity from locations 1 and 2 to locations 6, 7, and 8. The items must go through one of three intermediate locations (3, 4, or 5) before going to its final destination. There are associated costs in shipping the items as indicated in the following table. The Army would like to keep costs to a minimum.

i.

To / From	3	4	5	6	7	8	Supply
1	**50**	**62**	**93**	**-**	**-**	**-**	**70**
2	**17**	**54**	**67**	**-**	**-**	**-**	**80**
3	**-**	**-**	**-**	**67**	**25**	**77**	**-**
4	**-**	**-**	**-**	**35**	**38**	**60**	**-**
5	**-**	**-**	**-**	**47**	**42**	**58**	**-**
Demand	**-**	**-**	**-**	**30**	**70**	**50**	

The network would look similar to the generic one above except you would add one additional destination node. The problem formulation would look as follows:

Let x_{ij} = the amount shipped from node i to node j

Min Z= $50x_{13} + 62x_{14} + 93x_{15} + 17x_{23} + 54x_{24} + 67x_{25} + 67x_{36} + 25x_{37} + 77x_{38} + 35_{46} + 38x_{47} + 60x_{48} + 47x_{56} + 42x_{57} + 58x_{58}$

s.t.

Supply

$x_{13} + x_{14} + x_{15} \leq 70$

$x_{23} + x_{24} + x_{25} \leq 80$

Demand

$x_{36} + x_{46} + x_{56} \geq 30$

$x_{37} + x_{47} + x_{57} \geq 70$

$x_{38} + x_{48} + x_{58} \geq 50$

Transshipment

$x_{13} + x_{23} - x_{36} - x_{37} - x_{38} = 0$

$x_{14} + x_{24} - x_{46} - x_{47} - x_{48} = 0$

$x_{15} + x_{25} - x_{56} - x_{57} - x_{58} = 0$

$x_{ij} \geq 0$ for all *i* and *j*

The template solution would look as follows:

Linear Programming Model Templa																			
	1 to 3	1 to 4	1 to 5	2 to 3	2 to 4	2 to 5	3 to 6	3 to 7	3 to 8	4 to 6	4 to 7	4 to 8	5 to 6	5 to 7	5 to 8			
Dec var:	0	70	0	80	0	0	0	70	10	30	0	40	0	0	0			
Objective:																)BJ value:		
Cost	50	62	93	17	54	67	67	25	77	35	38	60	47	42	58	11670		
																Constrain		
Less than or equal to constraints:																t values:		
1 Supply	1	1	1													70	<=	70
2Supply				1	1	1										80	<=	80
Greater than or equal to constraints:																		
6Demand							1			1			1			30	>=	30
7Demand								1			1			1		70	>=	70
8Demand									1			1			1	50	>=	50
Equality constraints																		
3Transshipment	1			1			-1	-1	-1							0	=	0
4Transshipment		1			1					-1	-1	-1				0	=	0
5Transshipment			1			1							-1	-1	-1	0	=	0

The answer is therefore: Source 1 sends all supplies through node 4, source 2 sends all supplies though node 3, then node 3 sends 70 to destination 7 and 10 to 8, and node 4 sends 30 to destination 6 and 40 to 8. The total cost would be 11,670.

6. References:

a. Winston, Wayne L. and Albright, Christian S., *Practical Management Science*, 3rd ed., Mason, OH, South-Western Cengage Learning, 2007.

b. Hillier, Frederick S. and Lieberman, Gerald J, *Introduction to Operations Research*, 9th ed., Boston, McGraw Hill, 2010.

APPENDIX A
SUMMARY OF FORMULAS AND PROCEDURES
(Return to Table of Contents)

Section Three, Descriptive Statistics:

Mean (Average)

Population: $\mu = \frac{\sum x}{N}$ Sample: $\bar{x} = \frac{\sum x}{n}$

Median = Middle value of the ordered data.

Mode = Value that occurs the most.

Range = $x_{max} - x_{min}$

Variance

Population: $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$ Sample: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$

Standard Deviation

Population: $\sqrt{\sigma^2}$ Sample: $\sqrt{s^2}$

Section Four, Probability:

$$P(Event) = \frac{Successful\ Outcomes}{Total\ Possible\ Outcomes}$$

$P(A\ and\ B) = P(A \cap B) = P(A) \times P(B)$ for independent events

$P(A\ and\ B) = P(A \cap B) = P(A) \times P(B|A)$ for non independent events

$P(A\ or\ B) = P(A \cup B) = P(A) + P(B)$ for mutually exclusive events

$P(A\ or\ B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for non mutually exclusive events

$$P(A\ given\ B) = P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Section Five, Inferential Statistics:

Confidence Interval: $\bar{x} - E \leq \mu \leq \bar{x} + E \;\; where\; E = z_{\alpha/2}\frac{s}{\sqrt{n}}$

Steps for a Hypothesis Test:

- State a claim (hypothesis) about the population
- Select level of risk (α) and determine critical value (z)
- State the decision rule
- Take a representative sample to verify the claim
- Compute sample mean and standard deviation
- Compute z score for sample mean
- Compute p value for sample mean
- Compare z score with critical z (or p value with α)
- Reject or Fail to Reject the claim

Section Six, Regression:

$\boldsymbol{y = b_0 + b_1 x}$ **where:**

$$\boldsymbol{b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}} \quad \textbf{and} \quad \boldsymbol{b_0 = \frac{\sum y - b_1 \sum x}{n}}$$

$$\boldsymbol{r^2 = \frac{(\sum xy - n\bar{x}\bar{y})^2}{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$$

$$\boldsymbol{r = \frac{b_1}{|b_1|}\sqrt{r^2}}$$

$$\boldsymbol{s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}}$$

Section Seven, Decision Analysis:

Single Attribute Expected Value Payoff Matrix:

1. Construct a table with alternatives, states of nature, payoffs, and probabilities.
2. Determine whether the payoffs are costs or benefits.
3. Multiply each payoff within each state of nature by the probability of the state of nature.
4. Add the weighted payoffs across for each alternative.
5. Select the alternative with the best expected value (lowest cost or highest benefit).

Single Attribute Decision Tree:

1. Start with a decision node.
2. Place branches for each alternative.
3. Place a chance node after each branch.
4. Place a branch for each state of nature for each alternative.
5. Place probabilities for each state of each on each branch and payoffs after each branch.
6. Multiply each payoff by each probability.
7. Add the weighted payoffs for each chance node.
8. Select the alternative with the best value (low cost or high benefit). Trim off the other branches.

Sensitivity Analysis:

- Identify values that are uncertain or you think may change
- Estimate Extreme values (Highest and Lowest possible values)
- Recalculate Expected Value for each alternative at each extreme
- Check if decision changes at each value
- If decision does not change, then value is not sensitive
- If decision does change, find value at which you are indifferent (where expected value is the same for the alternatives)
- Determine if the change point is important (compare to original value)

Multi-Attribute Decision Analysis:

1. Construct a table with the alternatives, attributes, and payoffs.
2. Apply screening criteria. Satisficing; eliminate alternatives that do not meet requirements. Dominance; eliminate alternatives that are dominated by other alternatives.
3. Convert qualitative data by using a predetermined scale (1-5 scale for our purposes).
4. Scale values by using simple additive weighting (after converting all numbers will be between 0 and 1, and all will be benefits:

Maximize Attribute: $Rescaled\ Score = \frac{Attribute\ Value}{Largest\ Attribute\ Value}$

Minimize Attribute: $Rescaled\ Score = \frac{Smallest\ Attribute\ Value}{Attribute\ Value}$

5. Get attribute weights from management.
6. Multiply scaled values by weights for each attribute.
7. Add weighted values across each alternative.
8. Select the alternative with the best (largest) value.

Section Eight, Project Management:

Steps in constructing a Network:

1. Construct a table with activities, predecessors, and times.
2. Draw the network. Begin at a start node and activities from each predecessor.
3. Forward Pass. Begin at 0 with each activity that has no predecessor. Add activity time. Bring end time to activities that follow. Activities that have more than one predecessor are given largest time for their start.
4. Backward Pass. Begin with the end time for all activities that go into the finish node. Subtract activity time. Bring back latest start times to preceding activities. Activities that are predecessors to more than one activity will get the smallest time.
5. Slack Times. Calculate slack times by subtracting latest times from earliest times.

Critical Path. Activities with 0 slack times are on the critical path. Remember that the critical path begins at the start and connects all the way to the finish.

Section Nine, Math Programming:

1. Define the problem and recognize the type of math program (linear, transportation).
2. Create a diagram if helpful (i.e, transportation diagram)
3. Define the objective.
4. Define the decision variables.
5. Construct the objective function.
6. Construct the constraints.
7. Transfer the model into your math program model (i.e, excel template)

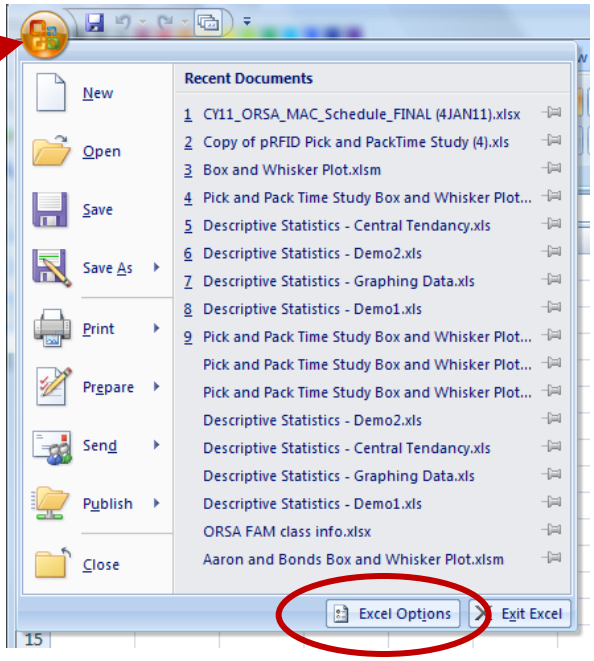
APPENDIX B

EXCEL ADD-INS AND TEMPLATES

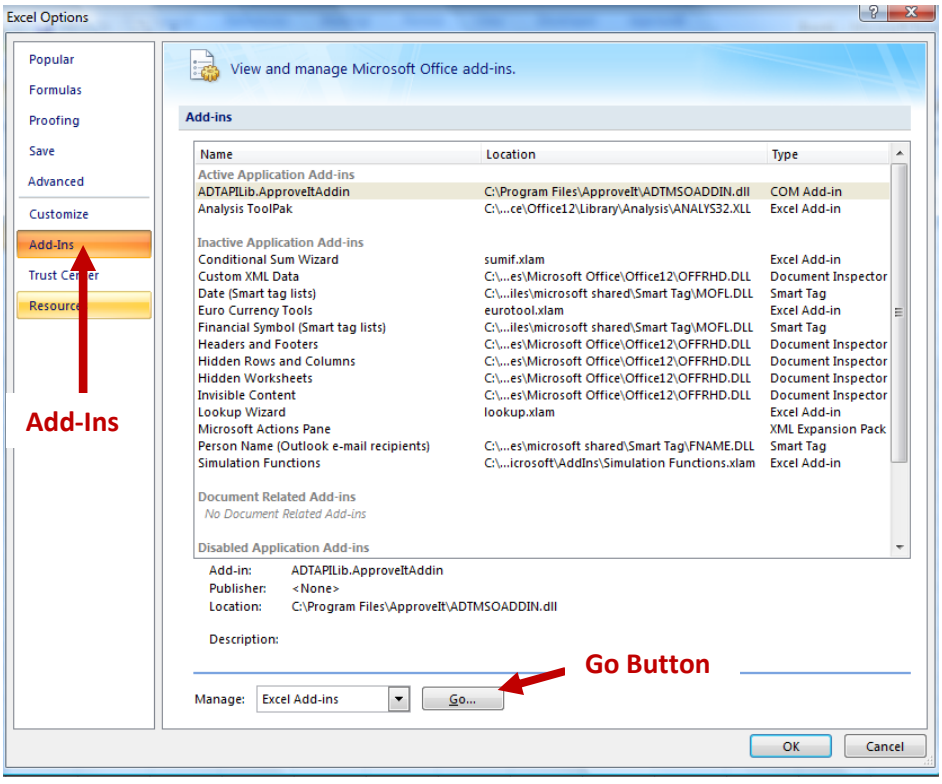
(Return to Table of Contents)

1. To install Excel add-ins follow the procedure below.

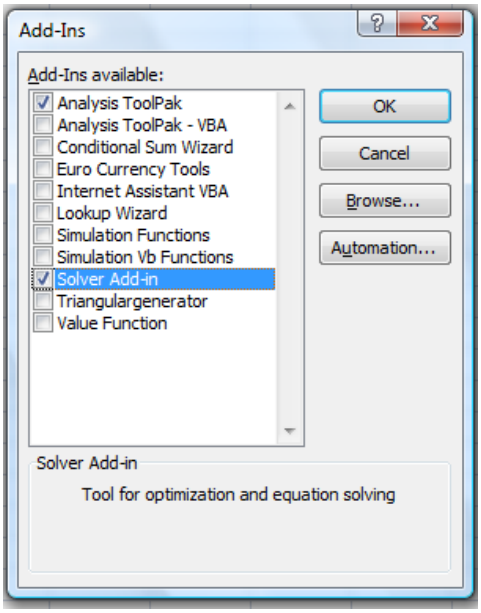
a. Click the **Excel Office Button** to reveal the menu shown below and then left **click** on the **Excel Options button (circled)**.



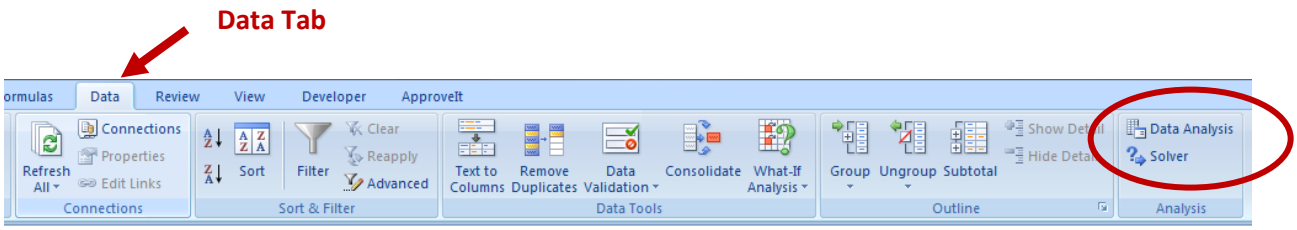
c. Clicking on the Excel Options button reveals the **Excel Options** menu **Highlight Add-Ins** and **click** on the **Go button**.



d. Clicking on the Go button reveals the Add-Ins menu. **Check** the **Analysis ToolPak** and **Solver Add**-in. Then **click OK** to install the selected Excel add-ins.



d. Once these add-ins have been installed you can find them to **clicking** on the **Data tab** to reveal the Data Ribbon shown below. The add-ins are **located at the far right** of the ribbon (**circled below**). To execute an add-in simply click on the desired add-in.

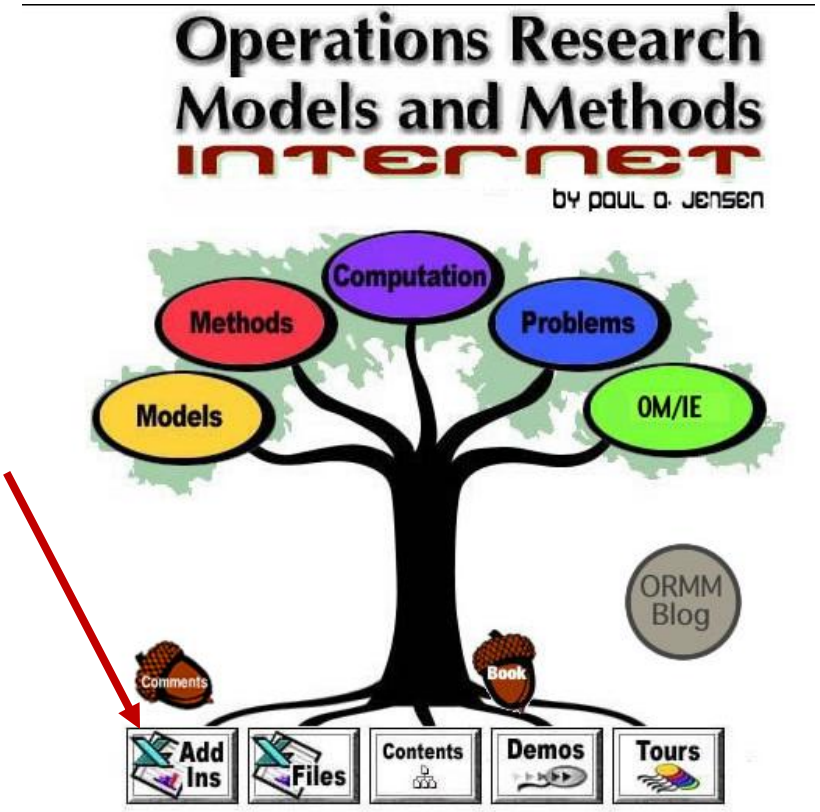


2. Dr. Jensen Excel Add-ins:

As discussed in several sections, Dr. Paul Jensen of the University of Texas has created several Analysis Add-ins for Excel. They are free to download and use. These add-ins are found at the following website:

http://www.ormm.net or http://www.me.utexas.edu/~jensen/ORMM/

When you reach his website, Click on the Add Ins icon on the bottom of the tree.



A listing of all the available add-ins will be displayed. You can download an individual add-in by clicking on it and saving, or you can go to the Jensen Library to download a zip file of all the add-ins.

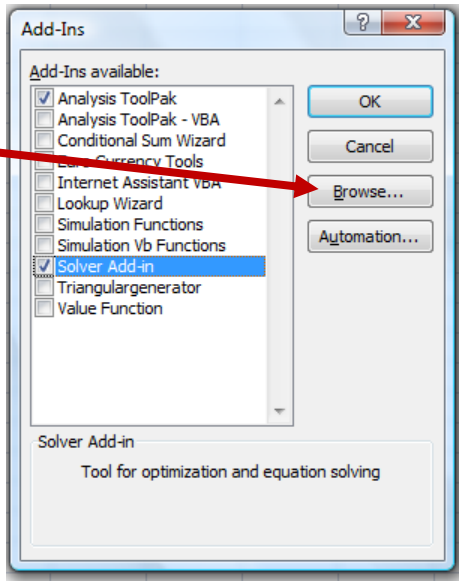
General Instructions for Add-ins **Document Revisions**
Jensen Library for Windows OS **Jensen Library for Mac OS**

Application Add-ins

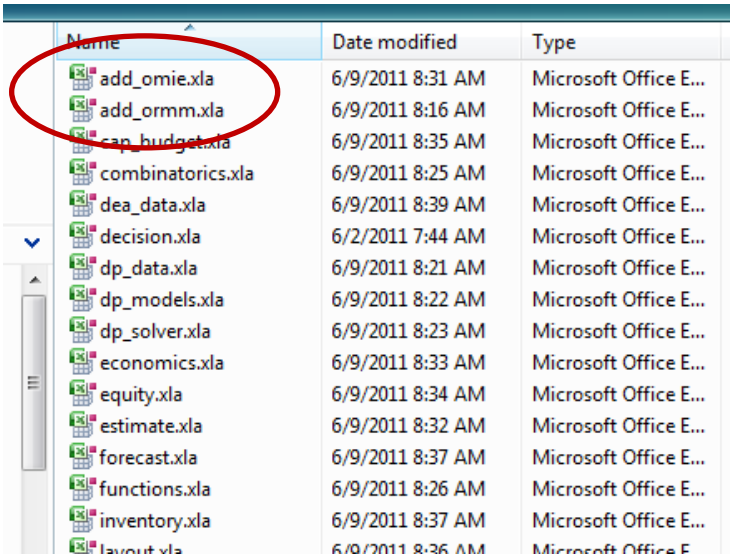
Add-ins hold macros that automate most of the features of model building and model solving. They must be installed rather than opened. Simply double clicking on the add-in file does not install it. Read the general instructions before attempting to use an add-in.

Operations Research Models and Methods	Operations Management /Industrial Engineering	Teach Operations Research
Instructions	Instructions	Instructions
Add ORMM	Add OMIE	Add Teach Add-in
MP Model Builder	Estimate	Teach Linear Programming
Math Programming Models	Investment Economics	Teach Transportation

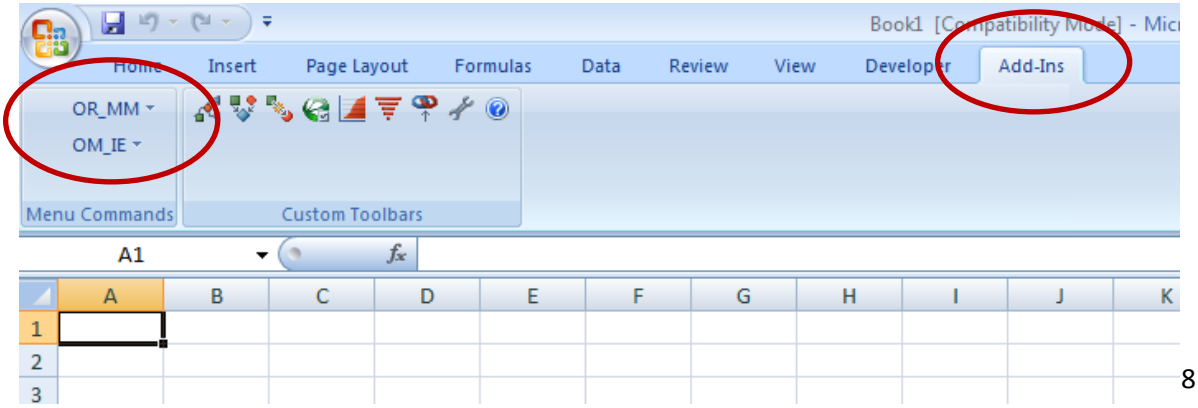
After you save the add-ins to your computer, you will have to activate them. Use the instructions in 1 above. When you get to the add-ins window**, Click on browse** and find the folder where you saved Jensen's add-ins.



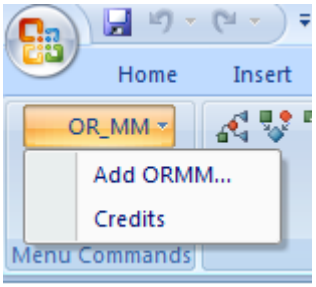
You only need to add two add-ins; add_omie.xla and add_ormm.xla.



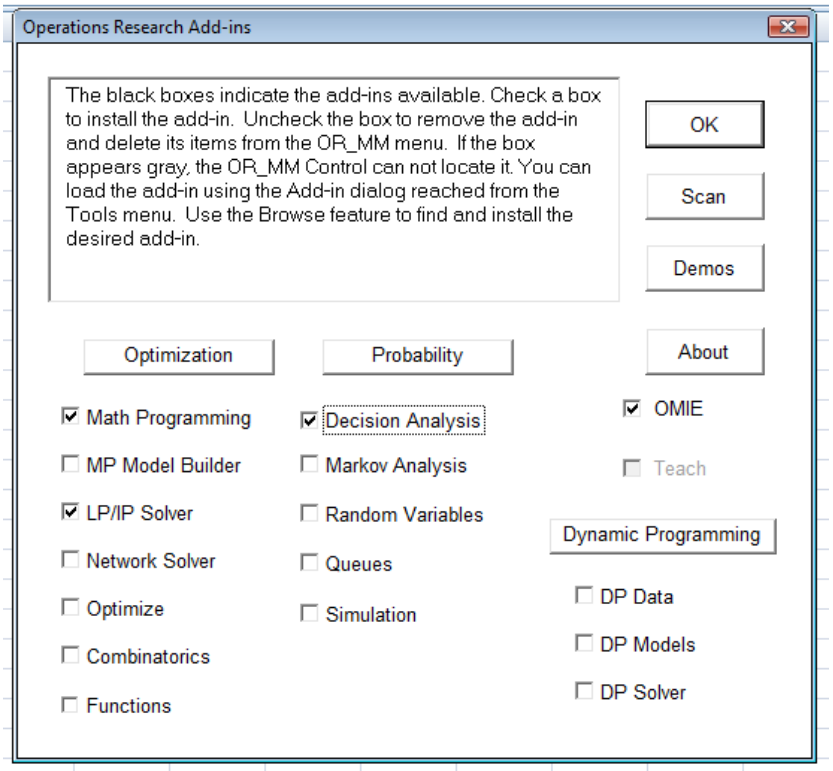
Once you activate those, in excel, go to the Add-Ins Tab and you will see OR_MM and OR_IE on the left side.



Click on one, i.e. OR_MM, then Click on Add ORMM…



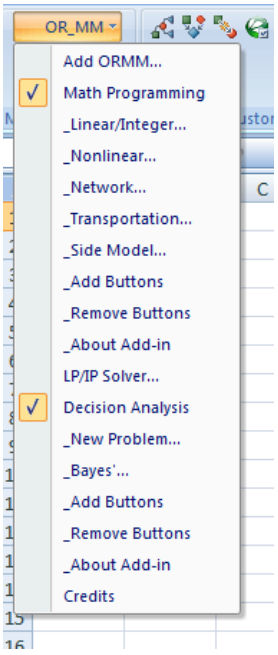
This will display a box where you can activate any of the add-ins you desire. Just place a check mark on the ones you want to use.



Clicking on OR_MM again will now display the activated add-ins which are now ready to use.

See the sections in this book for instructions on the use within the topics discussed.

See Jensen's website for further instructions.



3. Excel Templates.

a. Throughout the years, instructors within the ORSA committee of ALU have developed excel templates to assist students in class. These templates are useful for many different analyses and can easily be expanded or modified to fit individual needs.

b. Copies of these templates can be downloaded in the ORSA Net community of AKO found at the following web site:

https://forums.army.mil/secure/communitybrowser.aspx?id=436038

You must register to the community to download the files.

You can also receive copies by emailing or calling ALU.

c. The following templates are available:

1. Statistics:

Hypothesis Test Template

Box and Whisker Plot

Regression Prediction Template

2. Probability:

Distributions Template

3. Math Programming:

Math Programming Template

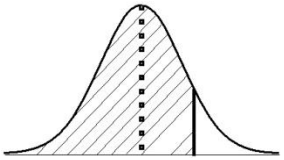
4. Decision Analysis:

Single Attribute Template

SAW Template (Multi-attribute simple additive weighting)

APPENDIX C
STANDARD NORMAL DISTRIBUTION

(Return to Table of Contents)



This table shows the cumulative values of the standard normal distribution from $-\infty$ to z.

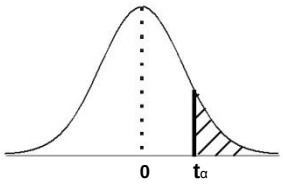
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004.	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0..0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0..0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1864	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4122	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Standard Normal Distribution Table (continued)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976

APPENDIX D
STUDENT t- DISTRIBTION

(Return to Table of Contents)



This table shows the critical values of the t-distribution

v	0.40	0.30	0.20	0.15	0.10	0.05	0.025
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.717	2.074
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.286	1.645	1.960

APPENDIX E
REFERENCES
(Return to Table of Contents)

1. A Glossary for Quantitative Analysis, Army Logistics Management College, 2000

2. Air Force Analyst's Handbook, On Understanding the Nature of Analysis, Feuchter, Office of Aerospace Studies, Air Force Materiel Command 2000

3. Decision-Focused Thinking, Dees, Dabkowski and Parnell, USMA, 2009

4. Decision Making in Systems Engineering and Management, Parnell, Driscoll, and Henderson, Wiley, John and Sons Inc., 2010

5. Decision Sciences for Logisticians, Army Logistics Management College, Thomson Learning Custom Publishing, 2001

6. Excel Add-Ins: www.me.utexas.edu/~jensen/ORMM/excel/project.html, Jensen, Paul

7. Guidelines for Army Analysts, How to Conduct an Analysis and Present the Results, Army Logistics Management College, 1989

8. Introduction to Operations Research, 9th ed., Hillier, Frederick S. and Lieberman, Gerald J, Boston, McGraw Hill, 2010.

9. Making Hard Decisions with DecisionTools, Clemen and Reilly, Duxbury, 2001

10. Mathematics for Managers Made Easy, Army Logistics Management College, 2001

11. Operations Research – Applications and Algorithms, 4th Edition, Winston, Brooks/Cole, 2004

12. Practical Management Science, 3rd ed., Winston, and Albright, South-Western Cengage Learning, 2007

13. Probability and Statistics for Engineers and Scientists, 8th ed., Walpole, Prentice Hall, 2007

14. Study Directors' Guide: A Practical Handbook for Planning, Preparing, and Executing a Study (TRAC-F-TM-09-023), TRADOC Analysis Center-Fort Leavenworth

15. The Operations Process, FM 5, Headquarters, Department of the Army, 2010

16. Value-Focused Thinking Using Multiple Objective Decision Analysis, Parnell, USMA